Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2008	**final**	**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Ten years of reform in primary mathematics education in England a review of effectiveness	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Ben Vollaard; Lila Rabinovich; Richard Bowman; Christian Stolk	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
RAND Corporation,1776 Main Street,Santa Monica,CA,90407	**TR-632-NAO**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Online access http://www.rand.org/pubs/technical_reports/TR632/

14. ABSTRACT

The UK National Audit Office (NAO) commissioned RAND Europe to look at the evidence showing the effectiveness of reform of mathematics teaching in primary schools in England. This study looked at government-sponsored evaluations of mathematics reform in England, independent evaluations of the educational outcomes of mathematics reforms in England, and the international evidence regarding some of the main components of government policy aimed at improving educational outcomes. The study presents the results of the evaluations and the international literature and discusses the robustness of their findings.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**68**	

This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

TECHNICAL REPORT

Ten years of reform in primary mathematics education in England

A review of effectiveness

Ben A. Vollaard, Lila Rabinovich, Richard Bowman, Christian van Stolk

Prepared for the National Audit Office

RAND EUROPE

The research described in this report was prepared for the National Audit Office.

Published 2008 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
Westbrook Centre, Milton Road, Cambridge CB4 1YG, United Kingdom
RAND URL: http://www.rand.org
RAND Europe URL: http://www.rand.org/randeurope
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

The UK National Audit Office (NAO) commissioned RAND Europe to look at the evidence showing the effectiveness of reform of mathematics teaching in primary schools in England. This study looked at government-sponsored evaluations of mathematics reform in England, independent evaluations of the educational outcomes of mathematics reforms in England, and the international evidence regarding some of the main components of government policy aimed at improving educational outcomes.

The study presents the results of the evaluations and the international literature and discusses the robustness of their findings. The study is divided into four main chapters: Chapter 2 gives an overview of the policy context; Chapter 3 discusses government-sponsored evaluations; Chapter 4 presents the findings of independent evaluations of mathematics (herein abbreviated to 'maths') outcomes in England and Chapter 5 discusses some of the international evidence on the effectiveness of specific components of government policy towards improving maths skills.

This report will be of potential use for policy-makers and evaluators in the area of maths education and Supreme Audit Institutions with an interest in 'value for money' work.

RAND Europe is an independent not-for-profit policy research organisation that aims to serve the public interest by improving policy-making and informing public debate. Its clients are European governments, institutions and firms with a need for rigorous, impartial, multidisciplinary analysis. This report has been peer-reviewed in accordance with RAND's quality assurance standards.

For more information about RAND Europe or this document, please contact:

Dr Christian van Stolk
RAND Europe
Westbrook Centre
Milton Road
Cambridge, CB4 1YG
Tel: +44 1223 353 329
Fax: +44 1223 358 845
Email: stolk@rand.org

Table of Contents

Executive Summary

The UK National Audit Office (NAO) commissioned RAND Europe to look at the evidence showing the effectiveness of the reform of mathematics (herein abbreviated to 'maths') teaching in primary schools in England. This study looked at government-sponsored evaluations of maths reform in England, independent evaluations of the educational outcomes of maths reforms in England, and the international evidence regarding some of the main components of government policy aimed at improving educational outcomes. The main findings were:

Maths teaching has undergone significant changes in England over the last two decades

Primary school education in England has undergone major reforms in the last 20 years, including the introduction of a National Curriculum at the end of the 1980s, national testing with league tables for schools in the 1990s, and detailed guidelines for teaching laid down in the National Numeracy Strategy shortly before the new millennium. These included new guidelines on the structure and content of maths lessons in primary schools in England. Additionally, spending per primary school pupil has greatly increased, with pupil to teacher ratios consistently falling over the last 10 years, according to the Survey of Teacher Numbers and Teacher Vacancies. To speed up progress in improving attainment, the Government has announced several initiatives, including a major review of primary maths teaching, focusing on subject knowledge of teachers (the 'Williams Review'), additional support for the lowest attaining primary school pupils ('Every Child Counts'), and pilots to provide personalised maths tuition ('Making Good Progress').

Government-initiated evaluations of reforms focus primarily on their implementation

Government-initiated evaluations tend to focus on the challenges and opportunities in the implementation of the reforms. Most of the evaluations discuss progress in the structure of mathematics education and teaching practices, rather than changes in pupils' educational attainment.

Nevertheless, this focus is still important because, when evaluating the effect of reforms in maths education, it does not suffice to look at changes in pupils' educational attainment. A crucial issue is how these changes came about, and implementation has a central role in assessing what worked, what did not work, and why. As the evaluations indicate, classroom practice is not always in line with official policy, at least not immediately, and reforms are not always uniformly implemented (Kyriacou, 2005: 173). In other words, the 'treatment administered' may differ from the 'intended treatment'. The relationship between reforms and pupils' attainment is mediated by the way in which changes are implemented, and the

resources – in terms of finances, human capital and time – that are available to put changes in place.

The impact of the reforms on pupil attainment are not robustly assessed in the government-initiated and independent evaluations

The lessons that can be learned from both government-initiated and independent evaluations are limited for the following reasons[1]:

- Evaluation and monitoring are often not considered at the design stage of the reform and reforms are mostly not piloted.
- When evaluating reforms that are uniformly 'prescribed' across the nation, it is difficult to include a control group.
- There is a general absence of evaluations that track performance over time, follow cohorts of pupils through the system, and methodically assess the factors that could shape this performance.
- Some of the reforms have been implemented only to a limited extent, precluding evaluation of some of the initiatives.
- The reforms often included a package of measures, which makes it difficult to see which (combination of) measures made the difference and which measures were not effective.
- It is often difficult in education evaluations to attribute impacts to individual factors such as teacher quality, class sizes, pedagogy, and types of assessment that influence educational outcomes.
- Many evaluations look solely at educational outcomes in terms of national test scores, without comparing progress in educational outcomes with independent and international tests on maths ability such as Trends in International Mathematics and Science Study (TIMSS)[2] and PISA (Programme for International Students Assessment)[3] or placing such test scores in the context of wider educational attainment in terms of further study or job outcomes.

Government-initiated evaluations show that while the prescribed format and structure for maths lessons have been adopted there is little evidence of 'deeper change' in teaching methods and improvements in pupil learning

In this study, we looked mostly at government evaluations of the Primary National Strategy (PNS) and National Numeracy Strategy (NNS).[4] These strategies were the most important reforms of maths education over the last ten years and they therefore merit special attention. The evaluations (Earl et al 2003; Ofsted 2002; Ofsted 2005; and Ofsted

[1] See also Chapter 5.

[2] See http://nces.ed.gov/timss/ (accessed June 2008).

[3] See http://www.pisa.oecd.org/pages/0,2987,en_32252351_32235731_1_1_1_1_1,00.html, (accessed June 2008).

[4] These reforms are described in Chapter 2.

2008) show that the changes in classroom practice tend to be in line with the NNS guidance, with most teachers using the format and structure of the prescribed daily mathematics lesson. Fundamental changes in methods of teaching – including 'assessment for learning' and 'stimulating mental calculation strategies' – have been slow in coming, even many years after the implementation. 'Deeper' change in teaching and learning has been hampered by poor subject knowledge, poor leadership, and limited understanding of the purposes of the NNS and PNS.

The research design of the evaluations does not allow for any definitive claims regarding the effectiveness of the reform in improving pupil attainment. The evaluations do not include control groups and do not control for other trends (such as the declining pupil to teacher ratio) that may have affected pupil test scores. As a result, the studies do not provide more than some hints as to what the impact of the reforms could have been. The results suggest that the impact of the NNS on pupils' attainment was mostly limited to the time period immediately after its implementation – when teachers' motivation was high and easy-to-implement changes in lesson structure were realised. In later years, gains in test scores were minimal.

High-stake testing[5] was found to result in considerable test preparation in the term leading up to the national assessments. How this affects pupils' learning is not clear, however.

Independent evaluations show a similar range of improvement in maths and attribute at least part of this to the NNS

The three independent evaluations (Brown et al 2003; Anghileri 2006; Basit 2003) show surprisingly similar overall gains in maths attainment up to four years after the implementation of the NNS. According to these evaluations, test scores improved by some 10 percent in the three to four years following the introduction of the NNS.

None of the studies provide hard evidence of the extent to which the gain in scores can be attributed to the NNS and later reforms. All studies are based on simple before/after comparisons, which leave the results open to alternative interpretations. Other national trends may well have played a role. However, the studies provide some indications that the NNS made at least part of the difference. Firstly, the effects of the NNS can be traced to relatively strong gains in test scores in areas that were emphasised in the NNS, including the number system and mental calculation. Secondly, the international comparative study, TIMSS, shows that gains in test scores over 1995–2003 only show up in year 5 and not in year 9, which suggests that the NNS made the difference, since the NNS is aimed at Reception through to year 6. Thirdly, lesson structure has evidently changed in line with the NNS and PNS, this being a necessary condition for the reforms to have had an effect. Nonetheless, like government-initiated evaluations, the independent evaluations find little evidence of more fundamental changes in teaching.

[5] High stakes tests are defined as those tests that carry serious consequences for students or educators. The consequences from standardised achievement tests range from decisions on passing and failing in subjects for pupils to rewards or punitive measures for schools and school districts (see e.g. http://www.eplc.org/mpearlman.html [accessed August 2008]).

The independent evaluations show that whole class teaching is of little benefit to low attaining pupils, and appears to favour boys over girls

All studies find a greater variance in test scores pre-reform compared with post-reform, with the lowest-performing quintile of pupils continuing to lag behind. The evaluations suggest that the greater variation could be the result of the emphasis on (interactive) whole class teaching, which was one of the requirements introduced by the NNS. The needs of low-attaining pupils are not fully met in a whole class setting; on average, they made almost no gain in attainment in maths. Some high-attaining pupils felt frustrated at their progress being held back; on average, their improvement was lower than that of average-performing pupils. Based on interviews and classroom observations, the evaluations suggest that interactive whole class teaching favours boys over girls, with girls perceiving it as competitive and unpleasant. Just like government-initiated evaluations (Earl *et al.* 2003 and Ofsted 2002), the independent evaluations find greater gains in scores for boys than for girls. TIMSS, however, does not confirm this result.

The study looked at international evidence from a number of interventions aimed at improving educational outcomes

This study examined the evidence in the literature on the effectiveness of a number of interventions, consisting of: class size reductions; formative assessments[6] (group teaching versus whole-class teaching); pedagogy (types of teaching); teacher quality; and standardised testing. These areas were identified in meetings between the NAO and the RAND Europe study team and were considered especially relevant for evaluating the past and present strategy of Government to improve maths outcomes. They are some identifying features of reforms of maths in the last decade.[7]

This overview aims to give important indications of effectiveness on the basis of meta-analyses and important field studies. However, this report does not aim to offer a comprehensive overview of all the evidence and the validity of each research approach. There are some important observations about the international evidence. A first observation in most studies is that attributing the effect on educational outcomes to specific factors and controlling for others remains difficult. Second, the meanings of concepts in education often overlap or are somewhat nebulous. There are overlaps between the use of formative assessment and teacher quality as a factor in improving educational outcomes and indeed between certain types of pedagogy, formative assessment and teacher quality. Moreover, the concept of pedagogy can include a wide range of teaching interventions. Therefore, a report like this has to take care not to confuse effects or overemphasise specific effects. Thirdly, the evidence regarding the impact of many interventions is often not specific to maths skills, but rather focuses on more general educational outcomes, effect sizes, or the equivalent of additional months of education in a given year. Thus, studies that look at the role of interventions such as reduction of class sizes, formative assessment, and personalised teaching in improving maths have to take care that the evidence is specific to the area of maths or indeed generalisable.

---

[6] Schools' use of assessment data to tailor the teaching to the pupil's need

[7] See for instance some frequently asked questions on the PNS, http://www.standards.dfes.gov.uk/primary/faqs/#236099 (accessed June 2008).

International literature suggests that teacher quality and the use of formative assessment are important factors in improving educational outcomes, while the effect of class size reductions is more moderate

Several studies point to teacher quality as the most important factor in affecting educational outcomes. This is an important observation, as teacher quality is often not the central component of government reforms. A study by Barber and Mourshed (2007) emphasises the importance of teacher quality and cites 1997 data from Tennessee on the difference in pupil outcome for pupils with low- and high-performing teachers over a three-year period. Other studies (Hedges et al 1994; Darling-Hammond 1998) have also found that teacher quality has a positive impact on pupil achievement. These studies raise important questions on which aspect of teacher quality (content knowledge, experience, pedagogy) contributes most to improvements in educational outcomes and also how governments can best support the development of teacher quality.

Formative assessments allow teachers to provide diagnostic feedback about their pupils' strengths and weaknesses and their instruction is guided by the results of assessments. Several meta-analyses (for instance Black and Wiliam 1998a and Black and Wiliam 1998b) show that formative assessments can have a positive effect on educational outcomes. The effect sizes[8] mentioned range between 0.4 and 1.8. In some cases the effect sizes studied were as large as in one-to-one instruction (Bloom 1984). From the evidence, the use of formative assessments appears an attractive proposition to policy-makers. However, the main question around the use of formative assessment is scalability. By definition formative assessment is a teacher-led intervention and studies have not looked at the effectiveness of the use of formative assessment when applied in a standardised way across an educational system.

There has been substantial research on the impact of the reduction of class size on educational outcomes. Though most studies highlight the importance of educational resources in promoting better educational outcomes, most agree that a reduction in class sizes will have only a moderate effect on pupil achievement (also distinguishing between educational outcomes and attainment). Moreover, the effect sizes differ for the various attainment groups and the intervention is potentially costly compared to other interventions for the same effect size. If class size reductions are to be part of the policy mix, these reductions should be directed at areas where the evidence shows the largest gains can be made, namely in classes containing low-achieving students and those in the first years of schooling.

There are important trade-offs in terms of cost-effectiveness, unintended outcomes and targeting of ability groups to consider when looking at the effectiveness of specific interventions

Some studies explicitly look at trade-offs between particular interventions. Several studies have found that small group instruction tends to have a greater effect on pupil achievement than whole class teaching (see Kulik and Kulik 1987; a meta-analysis in Davidson 1985; and Johnson et al 1995). However, the studies also found that improvements in pupil

[8] The effect size is the ratio of the average improvements in the test scores of pupils involved in an innovation over the range of scores for typical groups of pupils on these same tests (taken from Black and Wiliam 1998b).

achievement are often associated with particular teacher and pupil behaviours and interactions. Thus, the real underlying factors driving improvements in pupil achievement might not necessarily be the specific organisational arrangements in place but some of the characteristic behaviours found in interactions of smaller groups. These might also be replicated in whole class teaching.

A more recent study by Wiliam (2007) in the United Kingdom looked at trade-offs between formative assessment and reductions in class sizes in terms of their cost-effectiveness. He found that the use of formative assessment is about 20 times more cost-effective than the reduction of class sizes to achieve a similar gain in pupil achievement. Thus, the cost of interventions and the comparison of costs between interventions are areas where more research would be welcome and are important for policy-makers to consider when undertaking educational reforms.

In addition, certain interventions might have unintended consequences. For instance, class size reductions are associated with an increase of teachers within the educational system. Some observers have pointed to the effect of class size reductions on the quality of teachers, as new and not always the most qualified teachers are brought into the educational system. Such unintended effects could have consequences for pupil achievement.

Finally, certain interventions specifically affect pupil achievement in particular ability groups. Group teaching tends to have the largest effects in medium- and high-ability groups (respectively Webb 1991 and Good et al 1992). Though the evidence shows that class size reductions have a limited effect on educational outcomes, they do tend to be more effective in low-ability groups and with younger-aged pupils (Krueger 2000). Formative assessment tends to show the largest effects in low-ability pupils. These observations have to be taken into account when looking at wider trade-offs between interventions.

There is limited evidence regarding the impact of standardised testing

Standardised testing often provokes an emotive debate about its merits and its effects on educational outcomes. A positive aspect of such testing is that it makes it easier to track progress in educational outcomes. A negative aspect is that it could lead to the narrowing of the curriculum and 'teaching to the test', which might have a detrimental effect on numeracy skills. In poorly designed regimes some negative outcomes are noted from standardised testing, such as doubts over the validity and the robustness of reported test scores, and 'teaching to the test' is noted in some government-initiated evaluations in England. However, there is no systematic evidence that standardised testing has a limiting or negative effect on overall educational outcomes.

CHAPTER 1 Introduction

1.1 Background and objectives

The primary school experience lays the foundation for educational attitudes and progress in the later (educational) career. Poor educational outcomes may translate into inadequate mastery of basic maths skills in later life. Numeracy appears to have a significant independent effect on labour market opportunities, over and above the impact of years of education or educational credentials (Earl *et al.*, 2003: chapter 6). Based on UK data, Machin *et al.* (2001) found that individuals with better numeracy and literacy skills at age 16 will have higher earnings and higher rates of employment.

Primary school education in England has undergone major reforms in the last 20 years, including the introduction of a National Curriculum at the end of the 1980s, national testing with league tables for schools in the 1990s, and detailed guidelines for teaching laid down in the National Numeracy and Literacy Strategies shortly before the new millennium. Additionally, spending per primary school pupil has greatly increased, with pupil to teacher ratios consistently falling over the last 10 years according to the Survey of Teacher Numbers and Teacher Vacancies.[9] Although no strong claims can be made as to which reforms made the difference, the trend in young children's scores in statutory national tests is upwards. In 2007, less than a quarter of 11-year-olds had an insufficient grasp of the maths curriculum according to the minimum standard defined by the government, down from more than a half in 1995. Major concerns remain, however, especially in the area of maths. Based on hundreds of inspection visits to schools, Ofsted concludes that the quality of lessons in English is noticeably stronger than in maths in primary and secondary schools (Ofsted, 2008). Teaching and learning in maths is judged to be good in only half of the schools visited (Ofsted 2008). Upon leaving primary school, many pupils in England have insufficient maths skills to access the secondary curriculum.

To speed up progress in improving attainment, the Government has announced several initiatives, including a major review of primary maths teaching focusing on subject knowledge of teachers (the 'Williams Review'), additional support for the lowest-attaining

[9] The pupil to teacher ratio in primary schools in England has fallen from 23.4 in 1997 to 21.6 in 2008; the pupil to adult ratio in primary schools has fallen from 17.9 in 1997 to 12.0 in 2008.

primary school pupils ('Every Child Counts'), and pilot studies to provide personalised maths tuition ('Making Good Progress').

Against this background, the National Audit Office (NAO) is undertaking a value-for-money study of the educational outcomes for maths of primary school pupils in England. The NAO aims to assess whether government policies since the introduction of the National Numeracy Strategy in 1999 are having the desired effect on educational outcomes in maths, whether better outcomes also extend to children from disadvantaged backgrounds, and how schools are responding to the education challenge locally.

As part of this value-for-money study, the NAO has commissioned research reviewing the recent evidence on ways to improve maths education and attainment in primary schools. This report includes an assessment of the evidence regarding the effectiveness of reforms in England and a review of international evidence of the effects of different teaching modes and conditions on pupil attainment in maths.

1.2 Method

To analyse what the evidence tells about the effectiveness of the reforms pursued, we reviewed the literature evaluating the reforms since 1999.[10] These consisted of government-initiated evaluations of the major reforms, and independent evaluations. They are discussed in Chapters 3 and 4. Chapter 5 offers some observations on the validity and limitations of these evaluations.

In addition we also reviewed some of the international educational literature on aspects of primary maths education that are central to the reforms in England. This international literature, discussed in Chapter 6, provides further evidence which can give indications of the effectiveness of the reforms implemented in England. We include literature covering the following issues:

- Effectiveness of whole class teaching versus group or one-to-one teaching
- Effect of school class size on pupil attainment in maths
- Choices in pedagogy
- Effectiveness of formative assessment and personalised learning
- Impact of testing on pupil learning.

Given the wide scope of the reforms and large body of research on maths education on the one hand, and the limited time frame of this study on the other, RAND Europe and the

[10] Given the aim of this study, we limit the scope to reforms that are specifically aimed at improving primary maths education. Reforms that are broader in reach, affecting all of primary education, are outside the scope of this study. In addition, the consequences that maths education may have on other subjects – for example diminishing time on foundation subjects – are not within the scope of this study either (for instance, see Burgess, 2004 for a discussion on a 'broad and balanced' primary school curriculum).

NAO agreed that the study would review recent key publications rather than conduct a systematic review of the literature on all of these topics.

Because of the focus of this study on actual policies pursued in England, most of the literature reviewed was within the UK context. However, since the evidence base for the UK is limited, we also included evidence from other countries that have enacted similar reforms, such as the US.

CHAPTER 2 Reforms in primary mathematics education

2.1 Introduction

As discussed in the previous chapter, this study aims to assess the evidence regarding the effectiveness of the education reforms implemented in England since 1999 to speed up progress in pupil attainment in maths. This chapter provides an overview of these reforms.

2.2 National Numeracy Strategy (1999)

The National Numeracy Strategy (NNS) initiated by the Numeracy Task Force in 1996 constituted a large-scale reform in English primary schools. The key features of the NNS, which was implemented in all classes in primary schools at the start of the autumn term of 1999, concentrated around the structure and content of the lesson (Brown *et al.*, 2003):

Structure of maths teaching

- Three-part daily mathematics lessons of about 45 minutes in Key Stage 1 (year 1 and 2) and 50 to 60 minutes in Key Stage 2 (year 3 to 6). Lessons start with 5 to 10 minutes of whole class oral/mental arithmetic practice. This is followed by the main teaching activity of about 30 to 40 minutes involving direct interactive teaching of whole classes and groups. A plenary of 10 to 15 minutes rounds off the lesson.

- An emphasis on interactive whole class teaching. The lessons should be directed towards the class as a whole for most of the time. The aim is to incorporate both the highly and the less able children. Pupils needing extra attention should be included in the normal class as much as possible and be helped outside the plenary session. '[Very able children] can be stretched through differentiated group work, harder problems for homework, and extra challenges which they do towards the end of a unit of work when other pupils are doing consolidation exercises.' (DfEE, 1999: 20)

- A suggested week-by-week framework ('Framework for Teaching Mathematics from Reception[11] to Year 6') which introduced many skills at an earlier stage than previously. The framework is at a level of detail far exceeding that of the National Curriculum. Learning has been organised as a sequence of teaching objectives and milestones, contrasting with the up to then common focus on prescribed teaching tasks (Askew *et al.*, 2006).

Content of the lesson

- An increased emphasis on number and on calculation, especially mental strategies for calculation, including new methods of teaching number skills, a delayed introduction of written methods, and an encouragement for pupils to select from a repertoire of strategies. Calculators should not be used with pupils below year 5.

Implementation of the NNS was accompanied by a systematic and standardised national training programme, run locally by newly appointed local consultants and repeated by schools maths coordinators in all schools, using videos and transparencies to demonstrate 'best practices', with in-school support for schools perceived as needing it. Additionally, ambitious targets for minimum levels of achievement were set for improvement in test scores. The overall target was for at least 75 percent of 11-year olds to have attained level 4 of the National Curriculum tests for maths in 2002.[12]

2.3 Primary National Strategy (2003)

The Primary National Strategy (PNS) of 2003 brought together the National Literacy and the National Numeracy Strategies – with no substantive changes to the core of either of the strategies. The PNS exhorts schools to be more flexible and creative in managing the curriculum. It includes the following policies relevant to primary maths teaching (DfES, 2003: 29):

- Continuing funding for five-day courses to support teachers' own subject knowledge in maths
- Further guidance on key pedagogic practices to raise standards in maths, developing skills of questioning, demonstrating and modelling, and written presentational skills
- Promoting and supporting development of materials to help more able children and to tackle underachievement in maths of children with special educational needs
- Continuing to address how to teach effectively the aspects of maths that children find difficult, such as division, proportion and multi-step problems.

[11] Reception marks the transition from pre-school to primary education.

[12] National testing (the Standard Assessment Tests, SATs) was introduced at an earlier date. NNS set the following targets for each year: Year 1 level 1, and start on level 2; Year 2 consolidation of level 2, and start on level 3; Year 3 revision of level 2, but mainly level 3; Year 4 consolidation of level 3, and start on level 4; Year 5 revision of level 3, but mainly level 4; Year 6 consolidation of level 4, and start on level 5.

A further intervention in the day-to-day teaching of mathematics is the introduction of a model to reach children with learning difficulties. Intervention is seen in three stages or 'waves':

(1) For all pupils, as part of the lessons – the first principle remains that all pupils should be included as much as possible in the daily mathematics lesson

(2) For identified pupils, through focused work in small groups – for example, booster classes, springboard programmes, or other programmes linked to the NNS

(3) For individuals, through specific programmes.

The PNS performance target was for 85 percent of all primary school children to reach Level 4 at Key Stage 2.[13]

2.4 Every Child Matters (2004)

The policy initiative 'Every Child Matters' was aimed at providing chances to all children, including the low attaining pupils. 'Every Child Matters' did not include specific interventions in the area of primary maths education, and therefore is not described in detail here.

2.5 Renewed Primary Framework for Mathematics (2005)

In 2005, the NNS 'Framework for Teaching Mathematics' was renewed. Several general interventions outlined in the renewed Framework are important to mathematics education.

The renewed Framework reinforces the flexibility already introduced in the Primary National Strategy of 2003. The primary concern of the NNS was that the rigid use of the three-part daily maths lesson 'can act as a constraint on using the most appropriate organisation and structure to promote and develop children's learning' (DCSF, 2007: 8). 'The increased flexibility aims to encourage teachers in applying their teaching approach and pedagogy according to the needs of learners and the context of learning.' This idea of 'fitness for purpose' in pedagogy is at the core of the renewed Framework (*ibid.*: 11).

The structure of the renewed Framework supports 'multi-level curriculum planning'. If a child cannot work towards the same learning objective as the rest of the class, teachers may want to track back to an earlier objective. The Framework allows teachers to easily track

[13] Level 4 means that 'pupils are developing their own strategies for solving problems and are using these strategies both in working within mathematics and in applying mathematics to practical contexts. They present information and results in a clear and organised way. They search for a solution by trying out ideas of their own.' Taken from National Curriculum in Action, http://www.ncaction.org.uk/subjects/maths/levels.htm (accessed May 2008). The focus of teaching mathematics at key stage 2 may be on 'giving pupils opportunities to: use their awareness of space, shape and quantity in responding to the environment; compare, contrast and be aware of similarities and differences in shape, space and aspects of measurement; match and sort, selecting their own criteria; learn to count and use counting to find out 'how many?'; add and subtract in practical contexts; use numerals to represent amounts and respond appropriately to mathematical symbols; represent mathematical information in different forms and be able to make simple deductions.' Taken from the Qualification and Curriculum Authority website, http://www.qca.org.uk/qca_1857.aspx (accessed May 2008).

back and forward through a progression strand to locate earlier and later learning objectives (*ibid*: 14).

Additionally, the renewed Framework provides an even more detailed outline of what teaching maths should consist of in each year. As such, it provides a further specification of the NNS Framework.

2.6 Children's Plan (2007)

The Children's Plan sets out the government's objectives for improving children's wellbeing for the next ten years. It has a very broad focus. The aim of the plan is that every child is well prepared to go to secondary school, with no less than 90 percent achieving at or above the expected level in maths by age 11.

To prevent children from falling behind, the plan introduces a few new indicators, including the achievement gap between the lowest achieving 20 percent of children and the rest of the children at the end of the Early Years Foundation Stage; the achievement gap between pupils from low-income families who are eligible for free school meals and their peers achieving the expected level at Key Stages 2 and 4 (years 10 and 11); the proportion of pupils progressing by two levels in English and maths at each of Key Stages 2, 3 (years 7, 8 and 9) and 4; the proportion of children in care achieving Level 4 in English and maths at Key Stage 2; and the proportion of children in care achieving five A*–C GCSEs (or equivalent) at Key Stage 4 (DCSF, 2007: 56).

The plan also includes the provision of schools with better techniques for early identification and assessment of additional need, and support so that they can quickly refer to specialist services children who lag behind. These 'assessment for learning' techniques include the 'Making Good Progress'[14] tools for term-by-term tracking of progress that are currently being piloted at primary schools. The pilots provide up to ten hours of targeted one-to-one tuition in reading, writing and/or maths for 7–14-year-olds who are falling behind. The government plans to spend a further £1.2 billion over the next three years to adopt new teaching strategies including support for children with special educational needs and support for one-to-one tuition and small-group help.

2.7 Every Child Counts (2007)

Every Child Counts, set to start in 2010, will be aimed at children whose attainment in maths as 6-year-olds shows that they are failing to make expected progress for their age. Pupils will get intensive support each day from teachers, mostly provided one to one, but also through group work. It will reach approximately 30,000 6-year-old children by 2011 (DCSF, 2007: 69).

[14] They consist of assessment for learning and one-to-one tuition approaches among others. For an overview of Making Good Progress tools see http://www.teachernet.gov.uk/teachingandlearning/schoolstandards/mgppilot/ (accessed August 2008).

2.8 Conclusion

By far the most ambitious intervention in maths teaching in the last 10 years was the National Numeracy Strategy. Over the years, the NNS has been worked out in greater detail, new performance targets have been set, flexibility in applying the structure of the NNS has been emphasised, and more and more time and attention of teachers is focused on identifying and helping low-attaining pupils. Thus the emphasis on whole class teaching is increasingly supplemented by a range of pupil-specific interventions. This recent trend is interesting, since 'there was a perception that, prior to the introduction of the NNS, teachers in England had been trying too hard to differentiate mathematics tasks in order to meet the needs of individual pupils' (Askew et al., 2006: section 3.3). In the next chapters, we review the evidence on the effectiveness of the reforms in maths education described in this chapter.

When addressing this research question, the outcome will not be a simple judgement of how far actual policy practice is from an ideal practice as identified in the literature. As Brown *et al*. (1998) point out, the currently available evidence does not allow for simple recipes on 'what works and what doesn't' in the area of maths education. 'The complexity of the findings and of the possible interpretations suggests that ministerial desires for simply telling "what works" are unrealistic.' (*ibid.*) In most cases, the literature reveals that school, teachers, teaching organisations and teaching methods have a relatively small effect on numeracy attainment. Thus it is difficult to pick up consistent messages which relate to improved effectiveness – even when including evidence from similar reforms implemented in other countries.

Furthermore, the lessons that can be learned from both government-initiated and independent evaluations are limited, for the following reasons:

- It is hard to evaluate reforms that are uniformly 'prescribed' across the nation, making it difficult to include a control group
- Some of the reforms have been implemented only to a limited extent, precluding evaluation of some measures
- The reforms have often included a package of measures, which makes it difficult to see which (combination of) measures made the difference and which measures were not effective.

Clearly, that does not mean that no evaluation of the policies pursued is possible. Several studies evaluating the reforms have been published (such as Anghileri's 2006 study on pupils' calculating strategies for division before and after implementation of the National Numeracy Strategy), as well as studies analysing factors in maths education that are central to the reforms (such as Muijs and Reynolds, 2000 on whole class teaching). In this report, we review and summarise what is known and what is not known about the effectiveness of the reforms.

CHAPTER 3

Results from government-initiated evaluations

3.1 Introduction

To evaluate the effects of the National Numeracy Strategy, the Standards and Effectiveness Unit at the then Department for Education and Employment sponsored two parallel evaluations of its implementation, conducted by the Ontario Institute for the Study of Education (Earl *et al.*, 2003) and Ofsted (2002). Since then, Ofsted has published two further evaluations, in 2005 and 2008, which also cover later reforms, including the Primary National Strategy. Table 3-1 provides an overview of the government-initiated evaluations which we will discuss in this chapter. We do not limit our review to effects on pupils' test scores, but also include what changes have been brought about in the classroom.

Table 3-1: Overview of government-initiated evaluations of the reforms

Study	*Focus*	*Evaluation methods*
Earl *et al.*, 2003, 'Watching Learning 3. Final report'	NNS; 1998– 2002	Postal surveys to (head) teachers at 500 schools (2000 and 2002). Postal survey to numeracy consultants (2002). Repeated visits to 10 schools and interviews with numeracy managers and consultants from Local Education Authorities [LEAs] of these schools. Observations and interviews in 17 other schools and LEAs.
Ofsted, 2002, 'The NNS: the first three years 1999–2002'	NNS; 1999–2002	Inspection visits to representative sample of 300 schools in 1999; 200 schools in 2000–2002. Annual maths tests in years 3, 4, 5 in all 300 schools.
Ofsted, 2005, 'PNS. Evaluation of its impact'	PNS; 2004–2005	Inspection visits to 220 primary schools, survey across 47 LEAs, incl. meeting with primary strategy managers.
Ofsted, 2008, 'Evaluation of the PNS 2005–07'	PNS; 2005- 2007	Inspection visits to 85 primary schools, quota sampling based on Key Stage 2 test results, accompanying visits to 10 local authorities.

3.2 Evaluation by the Ontario Institute for Studies in Education

Researchers at the Ontario Institute for Studies in Education of the University of Toronto conducted a government-commissioned evaluation of the NNS. The evaluation covered the first years of the implementation of the NNS, up to July 2002. Results of the evaluation have been reported in two interim reports (Earl *et al.*, 2000 and Earl *et al.*, 2001) and one final report (Earl *et al.*, 2003). In our discussion, we focus on the third, final report.

3.2.1 Data

The primary source for the evaluation is surveys among and interviews with people directly involved in the implementation of the NNS. In 2000 and 2002, postal surveys were sent out to two samples of 500 schools, with questionnaires for headteachers and teachers. In 2002, a postal survey was also sent out to all numeracy consultants in LEAs. Additionally, the research team paid repeated site visits of 4 to 6 days to 10 selected schools (with various sizes, locations, pupil populations, levels of attainment) and their LEAs, including observations of mathematics lessons as well as interviews with headteachers and teachers. Finally, 17 other schools and LEAs were visited over the course of the evaluation.

3.2.2 Assessment of the evaluation

Mathematics teaching has changed in line with NNS guidance

According to Earl *et al.* (2003: 2), 'there is no question that the NNS has made substantial changes in primary education'. Effort and time spent on teaching of maths in primary schools appears, according to Earl *et al.*, to have increased as a result of the Strategy. Survey responses indicated that the majority of teachers and headteachers had implemented NNS to some extent in their classrooms by 2002. In addition, almost all schools had received some training for NNS, and teachers believed their own learning had been positively affected. The authors attributed the evident change to the 'high pressure, high support' approach of the reform, where performance targets and league tables were combined with support and increasing funding.

Schools indicated that the major shifts associated with the NNS had been (1) an improved range and balance of elements of maths being covered, (2) increased use of whole class teaching, (3) greater attention to the pace of the lessons, and (4) planning based on objectives rather than activities. Most teachers reported using the format and structure of the daily maths lesson. Earl *et al.* found that as teachers became more familiar and more comfortable with the framework and resources, many of them made adaptations to suit their pupils. The authors also reported that teachers and schools improved their capacity to use information from pupils' performance assessments to guide teaching. Greater use of assessment data offered a 'promising approach'; it was not common practice yet by 2002. Overall, the study concluded that 'there is considerable evidence that teaching has improved substantially since the NNS was first introduced' (*ibid.*: 3).

Gaps in subject knowledge and pedagogical understanding hamper 'deeper change'

When looking beyond the adoption of the structure and format of the daily maths lesson, the study found considerable variation in the extent to which teaching practice had changed. The authors argued that for many teachers, gaps or weaknesses in subject

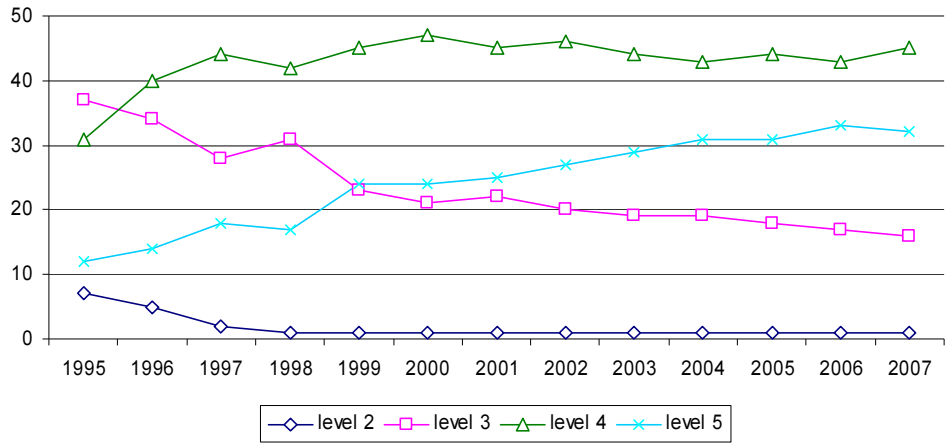
knowledge or pedagogical understanding limited the extent to which they were able to make full use of the framework and resources of the NNS. Many teachers noted that they did not feel secure about their subject knowledge and teaching of mathematics (*ibid.*: 47). For instance, not all teachers demonstrated an awareness of the different levels of understanding of each of their pupils, establishing curriculum targets for individuals while attending to the whole class and ensuring learning for all – one of the more fundamental changes the NNS was meant to bring about. The NNS provided teaching resources and training, but by July 2002 'many teachers have not yet had the sustained learning experiences necessary to develop a thorough understanding of the Strategy or of the best ways to teach mathematics' (*ibid.*: 6).

High-stake tests may undermine real progress, but evidence is unclear

Based on the data collected, the authors concluded that the national targets in the NNS skewed efforts in the direction of activities, 'some of them misinformed and counter-productive', that were intended to lead to an increase in the one highly publicised score (*ibid.*: 7). Many teachers acknowledged considerable test preparation, especially in the term leading up to the national assessments.

As the authors pointed out, the NNS also included countervailing forces. The emphasis on curricular targets and identifying the next appropriate learning objectives helped to broaden the focus beyond the scores on the Key Stage 2 tests (*ibid.*: 46). Another indication of a limited effect of the performance targets on re-allocation of teacher effort is that progress was not limited to those pupils who were below the threshold (*ibid.*: 46). The targets focused on the proportion of children achieving level 4 or higher, which may have led to a concentration of teachers' attention on children below the threshold – at the cost of children who were above the threshold. As Figure 3-1 shows, many more pupils have been achieving level 5 over the years, which does not support this hypothesis.

Figure 3-1: Substitution between level 3 and 5 drives change in maths test scores



Source: National Statistics

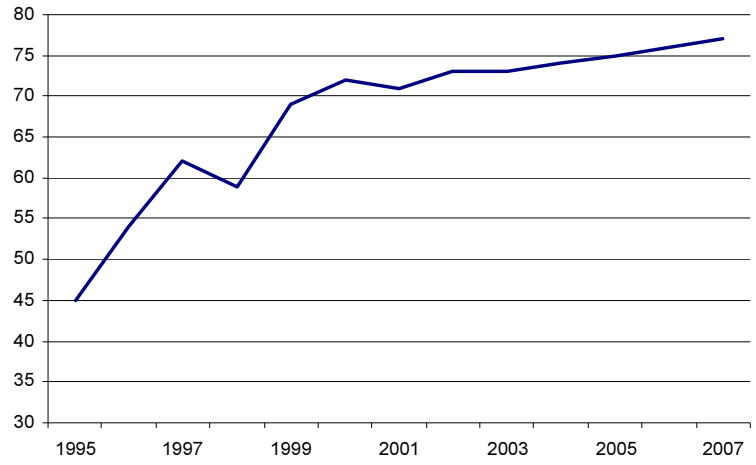
If inflation of scores due to test preparation had been a real issue, then the effect seems to have played out in the period prior to the implementation of the NNS. In 1995–1998, test scores rose sharply across the board, without an apparent reform that could drive such a quick and strong improvement in maths attainment. During these years, gains could have

come about because students and teachers became more familiar with the content and format of the standardised tests. In interviews, some head teachers and teachers expressed doubt over whether increases in test scores actually represented comparable increases in pupil learning. We come back to this issue in the next chapter.

Effect on pupil learning unclear

It is not clear from the evaluation how the changes in teaching affected pupil learning. The overall assessment of the evaluation is that 'increases in pupil learning have been considerable' (Earl et al, 2003: 128). As stated above, the largest increase in test scores occurred prior to the introduction of the NNS in 1999. The authors attributed the improvement in test scores immediately after the introduction of the NNS to greater motivation of teachers and others at the local level rather than to changes in their skills and knowledge (*ibid*.: 37). The results of more fundamental changes in teaching as a result of NNS were expected to materialise only after a few years. Maths results have changed only incrementally since 2002, however. Thus, any fundamental changes in teaching that the NNS brought about in the years since 2002 seem to have had a moderate impact on test scores (see Figure 3-2). Some head teachers and teachers are convinced, however, that 'pupil learning has improved considerably with the use of the NNS, with children showing increased understanding and skill in many aspects of mathematics' (*ibid*.: 3).

Figure 3-2: Percentage of pupils at level 4 or above, Key Stage 2 mathematics, England



Source: National Statistics

Value for money unclear

The evaluation does not draw any firm conclusions on value for money of the NNS. Earl et al. found that a relatively small amount of additional central expenditure (a 4.4 percent increase in annual expenditure for primary literacy and mathematics) has levered significant shifts in the use of schools' ongoing resources, such as teacher time and attention. Because of a lack of other comparative data, it is open to question whether the benefits exceed the costs and whether alternative ways of spending the public funds would have delivered better outcomes (*ibid*.: 125).

3.2.3 Discussion

The focus of this study is on evaluating whether and how the NNS guidelines were implemented. The study is highly qualitative in nature, with most of the evidence based on surveys and interviews with people directly involved in the implementation of the NNS. Given the research design and focus, no firm conclusions on the causal effect of the reform on pupil learning are possible.[15]

In spite of limited evidence regarding the impact of the NNS on pupil maths attainment, the evaluation does provide interesting insights into some of the challenges and opportunities of the reform. Firstly, the evaluation suggests that the reform appears to have placed mathematics as a top priority in classrooms across the country. Nevertheless, it also indicates that the needs for professional training necessary to effectively implement changes in teaching are greater than the reform had originally anticipated (although there appear to be discrepancies between the opinions of teachers, on the one hand, and experts, on the other, about the level of training and knowledge that teachers need). Secondly, the evaluation also suggests that the introduction of the reform itself may account for some of the improvements in student attainment since 1999 because it motivated both teachers and pupils. Thirdly, there is evidence that the Strategy leaders showed flexibility and willingness to adjust specific priorities and emphasis in response to emerging information about progress and challenges. For example the NNS produced materials and training when national assessment data showed that pupils had difficulties with problem solving in mathematics.

While it is difficult on the basis of the evaluation to make definitive assessments of the reform's impact on pupil attainment in maths, the report highlights a number of challenges and opportunities in the implementation of reforms like the NNS. Consideration of these challenges and opportunities can have longer-term effects on the ability of reforms such as the NNS to improve pupil attainment.

3.3 Ofsted's 2002 evaluation of the first three years of the NNS

In 2002, Ofsted published a review of the first three years of the NNS, based on inspection visits to hundreds of schools. The report summarises the standards attained by pupils, analyses the changes in teaching methods brought about by the strategy and suggests areas where further work is needed.

3.3.1 Data

The primary source of data was inspection visits to a nationally representative sample of 300 schools from 1999 to 2002. The sample was reduced to 200 schools in the second year of the implementation. The schools were visited at least once a year over the course of the evaluation. The inspectors observed the teaching of maths and held discussions with key personnel. They also inspected training and regularly met NNS maths consultants, their line managers in local education authorities (LEAs), and the regional directors of the

[15] Measuring the extent of implementation, and examining implementation relative to gains in student achievement could have provided a way of getting at the effect of the NNS on student attainment scores.

strategies. Evidence from inspections and from a telephone survey of 50 headteachers was also taken into account.

Additionally, Ofsted made use of the results of a testing programme for mathematics in year 3, 4 and 5 in all 300 schools. The testing programme was established by the Qualifications and Curriculum Authority; the data were collected and analysed by the National Foundation for Education Research (NFER).

3.3.2 Assessment of the evaluation

Greater teacher confidence and control

Ofsted (2002) concluded that the quality of teaching improved over the three years after the introduction of the NNS, although the biggest improvement was made in the first year of implementation. More teachers clearly stated the objectives of the lesson to pupils. Many teachers improved their knowledge and confidence in teaching maths with the help of the five-day training courses for teachers that are part of the NNS. According to the evaluation, the NNS provided teachers with the tools to 'regain control of the teaching of mathematics, rather than relying, as happened too often in the past, on pupils working their way through textbooks and worksheets' (*ibid.*: 25).

Plenary session weak; insufficient attention to mental calculation

Ofsted (2002) found that within the widely implemented three-part structure of the maths lesson, the oral and mental starter remained the best-taught element, with number being the main focus. The plenary session remained the weakest part of the daily maths lesson, with a lack of questioning and too much focus on the work of only one group of pupils, with the result that the rest of the class lost interest. Teaching improved over the period: the proportion of lessons where the plenary session was considered 'weak' in the judgement of school inspectors diminished from half in 1996 to one in six in 2002. Teachers gave insufficient attention to teaching mental calculation strategies and to linking mental strategies to written calculation methods – both essential elements of the NNS.

Greater pupil confidence

Ofsted (2002) also concluded that pupils' confidence, enjoyment of and involvement in maths had improved since the strategy began. Pupils responded positively to the routines and clear structure of the daily maths lesson and they were motivated by the direct teaching which it required. Many pupils understood their strengths and weaknesses in maths better, as well as the progress they were making (*ibid.*: 2).

Gains in test scores levelling off quickly; boys doing better than girls

The trend in (NFER) test scores over the period 1998–2000 was upwards, with the standardised score for year 5 pupils increasing from 99 in 1998, to 102.2 in 1999 and 104.1 in 2000 (Tymms, 2004: 486).[16] Since then, progress has levelled off, with a one percentage point increase in 2002 (Ofsted, 2002: 7). Boys achieved higher age-standardised scores than girls between 1999 and 2002 in all year groups.

[16] As discussed in Tymms (2004: 481), it is unclear why the standardised score for the first year, 1999, is equal to 99 rather than 100.

Three major concerns: weak assessment, poor subject knowledge, poor leadership

Three primary concerns emerge from the study: weak assessment practices, poor subject knowledge, and unsatisfactory leadership. Teachers had trouble with assessment practices, using the results of their own created assessments to adjust their teaching. Also, the study found that teachers were not confident in their teacher knowledge, and did not have a full grasp of all relevant mathematical concepts. Finally, the leadership and management of the strategy remained unsatisfactory in one in eight schools. This figure did not change over the three years of the implementation.

3.3.3 Discussion

Just like Earl *et al*. (2003), the Ofsted evaluation primarily focuses on implementation issues, rather than the effects of the NNS – although it notes that pupils' confidence in maths and teachers' confidence in their performance increased over the period. Statistics are mainly descriptive (e.g. frequency counts and means) and conclusions are mostly based on interviews with people directly involved in the implementation of the NNS. Based on the evaluation, it cannot be concluded how effective certain elements of the NNS are.

Nevertheless, the evidence provided by the evaluation on implementation issues is valuable, as it highlights some of the challenges facing improvements in the impact of the NNS and other reforms. The evaluation, for example, finds that problems in lessons planning, unsatisfactory teaching due to weak subject knowledge, and ineffective targeting of 'booster' provision continued to undermine improvements in maths teaching and attainment. The evaluation also highlights the important role played in implementation by the Local Education Authorities (LEAs) and strategy consultants.

3.4 Ofsted's 2005 evaluation of the Primary National Strategy

Following evaluation of the NNS, Ofsted (2005) evaluated the implementation of the Primary National Strategy (PNS) during its first years.

3.4.1 Data

From September 2004 until early 2005, school inspectors conducted a survey across 47 local authorities, including meetings with primary strategy managers and visits to 220 primary schools. The inspectors also monitored the work of a small number of primary strategy consultant leaders. A further 50 schools were visited as part of the evaluation of teaching and learning across the curriculum.

3.4.2 Assessment of the evaluation

Great variance in teaching methods between schools

PNS had its greatest impact through specific programmes, such as the Intensifying Support Programme (ISP), aimed at improving teaching based on the cycle of audit and setting targets, action and review. The reach of PNS was found to be limited, however. Schools were not always sufficiently aware of the range of support and guidance available to them. Ofsted (2005) concluded that although the NNS and the National Literacy Strategy had improved the teaching of maths and English, teaching in both subjects remained no better than satisfactory in one lesson in three. Great variance in the quality and consistency in

teaching between schools persisted. Moreover, in many schools, test scores rose in one year but fell the next, showing no consistent progress.

Longstanding weaknesses in leadership and 'assessment for learning' persisting

Longstanding weaknesses in leadership and management were not taken away by leadership programmes, Ofsted (2005) concluded. Despite support for teachers in developing assessment for learning, improvements were slow in coming and weaknesses remained. Schools' own use of assessment information often lacked rigour. The causes of underachievement were often not identified and remedied early enough, with too great an emphasis given to boosting achievement in years 5 and 6 at the expense of timely and effective intervention with younger pupils (*ibid.*: 4). The three waves of PNS intervention to support low-attaining pupils – adapting work within the classroom, small group support, and individual programmes of work – too frequently did not lead to success. Teachers were not sufficiently knowledgeable of the range of interventions available to them and which would be the most appropriate for different groups of pupils.

3.4.3 Discussion

Just like the previous evaluation by Ofsted, this study reviews which elements of the reform have been implemented – and highlights obstacles to implementation – rather than their effectiveness in terms of raising pupils' attainment in maths. As such, it is of limited use for assessing the effectiveness of the NNS and PNS.

The evaluation's focus on implementation, however, is informative on a number of issues central to the effectiveness of the strategy in the longer term. For example, like the previous Ofsted study, this one stresses the role of local authorities in providing the support necessary for schools to effectively implement changes. In spite of this support, building capacity within schools remains a challenge, reflected in fluctuating test results from year to year as opposed to sustainable year on year improvements. The evaluation states that teachers 'are not sufficiently knowledgeable of the range of interventions available to them and which would be the most appropriate for different groups of pupils' (Ofsted 2005: 5). The use of assessments for learning, for instance, continues to be one of the most problematic aspects of the reform.

3.5 Ofsted's 2008 evaluation of the Primary National Strategy

In its most recent report on primary education, Ofsted reviewed the effects of the Primary National Strategy during the years 2005–2007. It gave specific attention to the quality of transition from primary to secondary schools (which is outside the scope of this study) and to how well local authorities help schools to implement the Primary National Strategy.

3.5.1 Data

The findings are based on the inspection of 85 primary schools between autumn 2005 and spring 2007. The schools visited comprised a sample with equal proportions of schools with Key Stage 2 results above, in line with, and below the national average. Inspectors interviewed headteachers and key staff, talked to pupils, observed lessons and observed training and work in schools undertaken by local authority staff. In spring 2007, inspectors

visited 10 local authorities whose schools had been inspected during the previous two terms, evaluating the impact of the support provided for schools to implement the PNS.

3.5.2 Assessment of the evaluation

Lack of leadership and understanding of National Strategy hampers effectiveness

Ofsted concluded that the impact of the National Strategy on raising achievement was 'good' in around half of the schools inspected. It was most evident where there was strong leadership which ensured consistent approaches. In around one in ten schools, the impact was seriously limited by weak leadership, inaccurate self-evaluation and senior managers' low expectations of teachers. Teaching was least effective when teachers did not understand how the Strategy's recommended lesson structures could be used to help pupils learn. Inspectors found that the quality of lessons in English was noticeably stronger than in mathematics.

Poor 'assessment for learning' hampers effective intervention

Assessing pupils in lessons to ensure that learning was pitched at the right level continued to be the weakest aspect of teaching, according to Ofsted (2008). 'Any learning objective is meaningless unless properly linked to the appropriate developmental stage of (groups of) pupils.' (*ibid.*: 12) Too often, schools introduced intervention programmes without an accurate knowledge of pupils' weaknesses or used them as an alternative to good class teaching (*ibid.*: 15). Six previous evaluation reports identified the use of assessment as the weakest element of teaching and learning.

3.5.3 Discussion

As discussed before, the Ofsted evaluations are highly qualitative in nature, rich in context, but weak on evidence regarding what works and what doesn't. Ofsted (2008) concludes that the impact of the National Strategy on raising achievement was 'good' in around half of the schools inspected. The basis for this claim is qualitative evidence, including classroom observations. However, the report does not clearly explain how Ofsted (2008) arrived at this conclusion. Assuming that the report's conclusions regarding the reforms' effectiveness are correct, better leadership, greater understanding of the purposes of the Strategy, and better assessment are necessary to realise these benefits. Most importantly, and like the previous evaluations, this study emphasises that many teachers' lack of understanding of how the Strategy could help pupils learn hampered effective teaching.

3.6 Conclusion

NNS format and structure of daily maths lessons adopted, but no real change in teaching methods

Based on a review of the government-initiated evaluations, we conclude that while the NNS format and structure of daily mathematics lesson have been widely adopted, there is little evidence for 'deeper change' in teaching methods, even after many years, and the effects on pupils' learning seem limited.

The evaluations show that the NNS and later policy initiatives changed classroom practice. The studies primarily focus on implementation issues, rather than effectiveness of the

reforms: which elements of the reform were implemented;, how do teachers, other key personnel and pupils feel about it;, and what are obstacles to successful implementation. The changes in classroom practice tend to be in line with NNS guidance, with most teachers using the format and structure of the daily mathematics lesson. The plenary session is the weakest part of the lesson. More fundamental changes in methods of teaching – including 'assessment for learning' and stimulating mental calculation strategies – are slow in coming, even many years after the implementation. 'Deeper' change in teaching and learning is hampered by poor subject knowledge, poor leadership, and limited understanding of the purposes of the NNS and PNS.

NNS apparently positively associated with student test scores immediately after its implementation, with less impact in later years

The research design of the evaluations does not allow for any hard claims about the effectiveness of the reform on improving student attainment. The evaluations do not include control groups and do not control for other trends (such as the declining pupil to teacher ratio) that may have affected student test scores. As a result, the studies do not provide more than some hints as to what the impact of the reforms could have been. The results suggest that the impact of the NNS on pupils' attainment was mostly limited to the time immediately after its implementation – when teachers' motivation was high and easy-to-implement changes in lesson structure were realised. Nevertheless, the impact has been maintained, though in later years gains in test scores were minimal. This would suggest that more fundamental changes in teaching methods, if they were realised, had less impact on students' test scores in subsequent years.

Some evidence that strong incentives resulted in 'teaching to the test'

High-stake testing was found to result in considerable test preparation in the term leading up to the national assessments. How this affects pupils' learning is not clear, however. The performance target is defined as the percentage of students attaining at least level 4, which does not seem to influence negatively pupils who are already above the threshold. The percentage of pupils attaining level 5 in the Key Stage 2 test has been rising rapidly.

CHAPTER 4 Results from independent evaluations

4.1 Introduction

In this chapter, we review three independent evaluations of the effectiveness of the National Numeracy Strategy (NNS) of 1999, and the outcomes of an international comparative study into mathematics attainment, TIMSS. Later reforms have not been subjected to independent empirical evaluation yet. Table 4-1 provides an overview of the studies that we discuss in this chapter.

Table 4-1: Independent evaluations and international study

Study	*Focus*	*Evaluation methods*	*Main result*	*Exposure to NNS*
Leverhulme Numeracy Research Programme, Brown et al. 2003, How has the NNS affected attainment and teaching in year 4	NNS (overall); 1998 vs. 2002	Survey of year-4 pupil attainment, interviews with teachers, pupils	Test scores 4th year +10%*, spread +3%	3 years (3 out of 4 years of education)
Anghileri, 2006, A study of the impact of reform on students' written calculation methods	NNS (strategies for division); 1998 vs. 2003	275+ year-5 pupils' written responses to 8 division problems	Test scores 5th year +12%, spread +5%	4 years (4 out of 5 years of education)
Basit, 2003, Changing practice through policy: trainee teachers and the NNS	NNS (effect on teacher training); 1999/2000	Interviews with 30 final year trainee teachers at 1 university	Helpful for teacher trainees	1 year (1 out of 4 years of teacher training)
TIMSS, 2004	n.a.; 1995 vs. 2003	Maths tests for year 5 and year 9 pupils	Test scores 5th year +10%, spread +12%	4 years (4 out of 5 years of education)

Notes: spread denotes standard deviation * Test scores on number system items.

4.2 Leverhulme Numeracy Research Programme

The Leverhulme Numeracy Research Programme was a five-year study at King's College London during 1997–2002 funded by the Leverhulme Trust. The Nuffield Foundation provided a grant to extend the study for a further year in order to allow fuller data to be collected in 2001/02, with analysis and some follow-up in 2002/03. The study was not designed to evaluate the NNS, but given the timing of the data collection efforts (two years before and two years after the implementation of the NNS) it provided valuable information on the effects of the NNS (Brown *et al*., 2003).

4.2.1 Data

The core project entailed collecting comparable data on year-4 pupils in 1997/98 and in 2001/02. The children were tested towards the beginning and end of each school year. Test items were designed to assess conceptual understanding and cognitively based skills in numeracy. As with the NNS, the emphasis was on mental rather than written processes, and contextual as well as purely numerical items were included.

In the analysis, the only children included were those who completed the tests both at the start and end of year 4 from the 35 of the 40 schools involved which completed all four tests, giving samples of 1328 pupils for 1997/98 and 1291 pupils for 2001/02. The schools were selected by quota sampling to ensure a range of schools according to five variables (size, religious affiliation, socio-economic status of intake, attainment in national mathematics tests, and mathematical value added).

Additionally, classroom practice was observed, and classroom teachers, headteachers, mathematics coordinators and pupils interviewed. Such information is essential to interpret the measured change in pupils' attainment.

4.2.2 Assessing the evaluation

Table 4-2 provides an overview of the changes in maths attainment after implementation of the NNS. Below, we discuss the interpretation of these changes put forward by Brown *et al*. (2003).

Average scores are up; improvement associated with curriculum change

Based on a comparison of average test scores, Brown *et al*. (2003) concluded that the NNS resulted in a 'small change in overall attainment'. Scores in year 4 improved by about 3 percentage points. The increase is statistically significant (at the 1 percent level) and equivalent to 2.4 months' development. Two-thirds of the schools had higher results in 2002 than in 1998; the remaining third had lower results. Brown *et al*. (2003) suggest curriculum change as the major factor, with a resulting improvement in pupils' facility on items relating to numbers and the number system and place value.

Girls stay behind; girls dislike whole class questioning

The gap between girls and boys widened after the introduction of the NNS, although both groups made small gains. Based on classroom observations, Brown *et al*. (2003) suggest that the public nature of the whole class questioning favoured both more competitive children and those who preferred oral to written work; in each case, these were more likely to be boys. In interviews, low-attaining girls expressed dislike for the mental/oral work

since they were worried that they would not give the correct answer and then the rest of the class would consider them 'stupid' (*ibid.*: 664).

Table 4-2: Student test scores in year 4 before and after introduction of the NNS, by sex, attainment level, level of deprivation*, area of curriculum, standard deviation

	June 1998	*June 2002*	*% point change*
Mean student % score	62	65	+ 3
Boys	62	66	+ 4
Girls	61	63	+ 2
Difference	1	3	+ 2
Best performing 25%	82	85	+ 3
Average performing 50%	63	67	+ 4
Lowest performing 25%	39	40	+ 1
Difference best vs. lowest performing	43	45	+ 2
Least deprived 25%	68	71	+ 3
Middle 50%	61	65	+ 4
Most deprived 25%	56	61	+ 5
Difference least and most deprived	12	10	– 2
Number system	63	69	+ 6
Addition/subtraction	60	65	+ 4
Fractions/decimals/ratio	41	42	+ 1
Real-life problems	40	39	- 1
Multiplication/division	72	68	- 4
Standard deviation in scores	17.1	17.6	+ 0.5

Source: Brown et al. (2003)

Note: * Based on the Townsend index of social deprivation for children's postcode.

Lowest-attaining pupils stay behind; deriving little benefit from whole class teaching

While the middle 50 percent and the best performing 25 percent showed gains in attainment, the 25 percent lowest-performing students made almost no gain in mathematical attainment and the lowest-attaining 5 percent of students saw a small decline in scores (not shown).

Both observations of lessons and interviews with children suggest that low-attaining pupils derive little benefit from the whole class teaching episodes, and the topic of the lesson does not always correspond to their areas of greatest need (*ibid.*: 662). In interviews, teachers confirmed that lower-attaining pupils were not able to participate satisfactorily and confidently in whole class teaching episodes. They felt that the needs of this group were not being fully met, although schools were struggling to do this through the use of

teaching assistants. Additionally, some high-attaining pupils also expressed their frustration at their progress being held back by the whole class teaching emphasis, which tends to be pitched at the needs of the middle of the group.

Pupils from deprived neighbourhoods made the greatest gains

The most deprived 25 percent of pupils made greater gains between 1998 and 2002 than the least deprived. Pupils from deprived backgrounds seemed to benefit from higher expectations and a more uniform curriculum (*ibid.*: 665).

Gains in number system and addition/subtraction greatest; result of change in focus

Changes in attainment differed across the numeracy curriculum. Improvements in the number system and addition/subtraction were not matched in multiplication/division and real-life maths problems.

Brown *et al.* (2003) attributed this imbalance to a tendency to focus on addition and subtraction at the expense of multiplication and division. Pupils' approaches to real-life problem solving and the using and applying of mathematics may also have suffered through a decreased focus on these areas.

No evidence of encouragement of pupils' strategic thinking

Three years into the implementation of the NNS, observation data showed limited evidence of what the Strategy had recommended in terms of the encouragement of strategic thinking (*ibid.*: 668). The NNS stresses the importance of pupils not only developing a 'repertoire' of mental and written calculation strategies from the earliest years but more importantly an ability to select between these according to the size of the numbers and the purposes of the calculation. Brown *et al.* (2003) did not find an increase in teaching that would promote this strategic thinking. Their analysis of the post-Strategy lessons shows that pupils were provided with more opportunities to explain their methods. They found little evidence, post-Strategy, of pupils actually discussing different methods and looking at their strengths and weaknesses in applying them to different calculations.

Brown *et al.* (2003) suggested that a move to more strategic ways of working was one of the more difficult aspects of the NNS for teachers firstly to understand and secondly to implement. The more immediately understandable aspects of a reform are likely to be implemented first; in this case, those of the structure of the lesson, the changes in planning, and the teaching of a range of methods illustrated in the Framework document. Those aspects that cannot be accommodated through adapting existing practice or require a deeper understanding of mathematical principles require further professional support. Some headteachers and maths coordinators interviewed in 2003 had identified this move from procedural to strategic ways of working as their next priority.

Box 4-1: Change in organisation and structure of mathematics lessons

In a related study within the Leverhulme Numeracy Research Programme, Askew et al. (2006) surveyed how the introduction of the NNS changed teaching methods. A teacher survey (with 62 participating teachers in 1997/98 and 55 in 2001/02) was combined with interviews with teachers and lesson observations.

Interactive whole class teaching. The percentage of time devoted to whole class teaching increased from 46 percent in 1998 to 63 percent in 2002 – at the cost of individual work. Teachers in 2002 felt that there was more discussion between teachers and pupils and that pupils were explaining their methods more frequently. However, as Brown et al. (2006) pointed out, the greater use of whole class teaching started to take place well before the introduction of NNS. In 1998, 80 percent of year 5 teachers taught mathematics to the whole class for most days, up from 11 percent in 1994. The remainder were regularly teaching the whole class together and ensuring that they were working on the same topic.

Three-part lesson. Few teachers in 1998 structured their lesson in three parts, as suggested by the NNS, whereas in 2002 almost all lessons had three parts. The time allocation for the main part of the mathematics lesson and the total lesson length were very close to the times recommended by the Strategy.

Use of the Framework. The Framework was being used by two-thirds of year 4 teachers at least weekly for planning their teaching in 2002, with the remainder mainly using commercial schemes.

Objectives-based planning. The move from activities-led planning to objectives-led planning was mentioned in all five schools in 2003.

4.2.3 Discussion

By 2002, it was still early to judge the effectiveness of the NNS. The year 4 students that were the subjects of the study had only been exposed to the NNS later in their school career. Moreover, as already noted in the evaluations reviewed in the previous chapter, Brown *et al.* (2003) argue that it takes a longer period before the harder to implement parts of the NNS, such as encouragement of pupils' strategic thinking, are picked up. Therefore, this evaluation provides a limited picture of the effects of the NNS, mainly focusing on the effects of organisation of lessons and resources used.

Brown *et al.* (2003) argue that the small positive change in student test scores after the implementation of the NNS was mainly the result of the change in the curriculum. This could either be the result of the fact that this factor did change and other factors such as pedagogy did not *or* the fact that curriculum change matters and factors like pedagogy does not. Clearly, these alternative explanations of the same finding have very different policy implications. Studies with stronger research designs are necessary to see which hypothesis is supported by the data. Given the weak research design underlying the evaluation, even the finding that curriculum change is the major factor needs to be interpreted with caution. The researchers do not discuss how other factors may have affected the test scores. In addition, conclusions as to what is driving the changes in test scores are mostly based on qualitative evidence (interviews and classroom observations). However, their conclusions

are supported by the fact that gains are highest in the areas that the NNS focuses on (number system).

The finding that boys do better than girls after the introduction of the NNS confirms the conclusion of Ofsted (2002).

4.3 Anghileri 2006

Shortly after the implementation of the NNS, Anghileri (2006) replicated a study on year 5 pupils' calculating strategies for division that was conducted in 1998 (Anghileri 2000, 2001). In line with the guidance of the NNS, by year 5, most pupils (9 and 10 year olds) can be expected to have established good understanding of numbers and a range of mental strategies for calculating.

4.3.1 Data

Written responses to eight comparable division problems were collected for 275 pupils in 1998 and 308 pupils in 2003. Both studies used substantially the same items and relied on pupils from the same set of schools (1 out of the 10 schools included changed). Creating a nationally representative sample is not stated as an explicit aim.

4.3.2 Results

Average scores are up; greater variation in scores

After implementation of the NNS, average scores increased by 12 percent, but the standard deviation also increased by 5 percent. Brown *et al.* (2003) found a much smaller gain for multiplication/division items: 6 percent. A somewhat higher result was to be expected, however, since Anghileri (2006) included pupils who were one year older and who had been exposed to the NNS for an additional year compared to Brown *et al.* (2003).

Little change in calculation strategies; great variation across schools

Few pupils used a mental strategy (i.e. they gave an answer but showed no working) (Anghileri, 2006: 374). Although the NNS introduced emphasis on more flexibility in strategy choice according to the numbers involved and the context, the number of different strategies used for the division problems varied little from the 1998 results. The results showed a shift from extensive use of the traditional algorithm in 1998 to more use of informal methods in 2003 and different written methods.

Pupils' responses indicated that teachers had interpreted the 'Framework for Teaching Mathematics' in different ways. Additionally, although methods to develop structured written methods were illustrated in the Framework, it appeared that some teachers, several years after implementation of the NNS, were not providing pupils with the necessary support. This suggested that the widespread distribution of guidance was not readily followed by equally widespread implementation. 'Documentation alone is not sufficient to develop teaching approaches, and further professional engagement with curriculum reform is necessary to help teachers understand the deeper purposes of change.' (*ibid.*)

Girls stay behind

Implementation of the NNS coincided with the emergence of a gap in mean scores between boys and girls. In line with the findings of Ofsted (2002) and Brown *et al.* (2003), gains were more pronounced for boys than for girls. By 2003, the gap was statistically significant at the 1 percent level. Boys and girls also differed in their calculation strategies, with boys using informal and mental strategies more often than girls. The difference between boys and girls was not uniform across all the schools and in some schools the girls did better than the boys.

4.3.3 Discussion

As in the case of Brown *et al.* (2003), the conclusions are based on a simple before-after design. Therefore, it is not clear whether the measured gains in test scores can be fully attributed to the NNS. But again, effects of the NNS can be traced to changes in the types of answers, lending support to the hypothesis that the NNS did make a change. The internal validity of the study is strengthened by using (almost) the same set of schools in 1998 and 2003. External validity is weak given the small sample, which does not seem to be representative of the wider population of schools and pupils in England. Importantly, Anghileri (2006) provides empirical support for the claim of other studies, including Brown *et al.* (2003), that teaching methods have not changed much.

4.4 Basit 2003

To see how the NNS affected teacher training, Basit (2003) analysed how its introduction was perceived by trainee teachers. The research aimed to find out: what the effects of a top-down initiative were on a group of primary teachers; whether they found the NNS helpful; and how it affected their understanding of teaching primary mathematics.

4.4.1 Data

Thirty final year primary B.Ed. students at one university were interviewed in depth. This group of trainees was chosen because they had done three years of training without the NNS being in place and then a fourth year of school experience to which the NNS applied. Thus the study was carried out during the first year of the NNS (1999/2000). The sample included those who readily agreed to take part in the study (roughly half of the students who were approached).

4.4.2 Results

Based on the interviews, Basit concluded that student teachers saw the NNS as a helpful framework for developing their professional expertise in an area where they often had experienced some anxiety. The NNS helped them in planning and developing maths teaching programmes and activities within them, which was surprising for a top-down policy. The biggest concern identified was that rigid time management (based on NNS guidelines) could be problematic for lower-attaining pupils if they were being moved on to the next topic before adequately grasping the current topic.

4.4.3 Discussion

The study provides interesting perceptions of the effects of the introduction of the NNS, but lacks external validity: it is not clear to what extent the findings can be generalised to the whole population of trainee teachers. The views of these teacher trainees are not necessarily representative of either their direct colleagues (e.g. other young students from British descent) or the whole group. Additionally, since students were going to be exposed to new material in their final year anyway, they only had a limited idea of how NNS changed their teacher training. The most important finding is that the reforms do not seem to have met with strong resistance, which would have lowered their effectiveness.

4.5 TIMSS

The Trends in International Mathematics and Science Study (TIMSS) is an international assessment of the mathematics and science knowledge of fourth- and eighth-grade students around the world. The study was developed to allow participating nations to compare students' educational achievement across borders. TIMSS was first administered in 1995, and every four years thereafter. The most recent results available are for 2003. The results of the 2007 survey, in which some 60 countries participated, will be available by the end of 2008.

4.5.1 Data

TIMSS consists of an assessment of mathematics and science, as well as student, teacher, and school questionnaires.[17] TIMSS includes mathematics achievement of year 5 pupils for 1995 and 2003, and year 9 pupils for 1995, 1999, 2003.[18]

4.5.2 Results

Figure 4.1 shows the development over 1995–2003, with the US as the comparison country. While scores for year 9 pupils remained stable, scores of year 5 pupils greatly improved in England, indicating that the NNS did have a clear impact on maths attainment. After all, the NNS was focusing on pupils up to year 6. The year 5 pupils whose 2003 test scores were reported had been exposed to the NNS from their second year of education. Crucially, the year 9 pupils had only received exposure to the NNS for one year (as year 6 pupils in 1999/2000). As such, TIMSS allows for a difference-in-difference analysis, comparing improvement in scores of pupils who did receive NNS-style teaching (treatment group, the year 5 pupils) and pupils who did not (control group, the year 9 pupils). Based on this analysis, gains in test scores for year 5 pupils did not seem to be the result of wider changes in education that would have equally affected year 9 pupils.

---

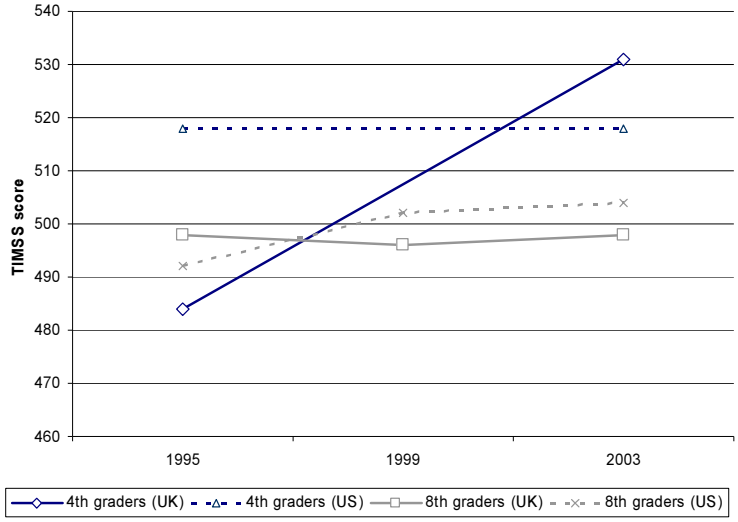
[17] The student questionnaires are designed to collect information on students' backgrounds, attitudes and beliefs related to schooling and learning, and information about their classroom experiences, among many other topics. The teacher and school questionnaires ask about class scheduling, maths and science content coverage, school policies, teachers' educational backgrounds and preparation, among many other topics.

[18] The 2003 scores for year 9 pupils did not meet the minimum sampling requirement, but can be expected to be representative nonetheless (see Ruddock *et al.*, 2004: 5 for a discussion).

The scores of year 5 pupils improved by 10 percent. This gain in test scores was substantial: it was much larger than the average change for similar countries (Ruddock *et al.*, 2004: 8). The increase for items assessing number, rather than other aspects of maths was higher (*ibid.*: 10), indicating that the NNS may have made the difference. While both year 5 and year 9 pupils trailed behind the US in 1995, English year 5 pupils passed their American counterparts in 2003.

In line with the other studies, the spread in scores increased (by 12 percent). Although variation increased, the survey did not support claims that there was a long tail of underachievement (Ruddock *et al.*, 2004: 13). The TIMSS results did not confirm the finding of Ofsted (2002), Brown et al. (2003) and Anghileri (2006) that the performance gap between boys and girls was growing. Scores of both genders increased significantly, with girls actually doing somewhat better than boys.

Figure 4-1: TIMSS maths scores in the UK and the US, 1995–2003



Source: TIMSS

4.5.3 Discussion

Similar to Anghileri (2006), the TIMSS data includes pupils that have been exposed to the NNS during most of their educational career thus far (four out of five years). Moreover, the results are stronger: they still stand when subtracting the change for year 9 pupils unaffected by the implementation of the NNS. As stated in the introductory chapter, comparisons of test scores alone give a limited picture of the effects of reform. Other factors could have affected test scores during this period – although previous research suggests that education is a major explanatory factor of differences in TIMSS test scores across countries (Harris et al. 1997).[19] Given the infrequent and no-stake nature of the TIMSS, 'teaching to the test' is of little concern. Also, the NNS focus on the number

[19] Based on England's relatively poor performance in maths but good performance in science, Harris et al. (1997) suggest that educational, rather than socio-cultural, factors are driving poor maths achievement of English pupils.

system and mental calculation is reflected in the TIMSS, since the tests were designed to reflect the maths curricular goals of the specific country.[20]

4.6 Conclusion

Based on a review of the three independent evaluations and one international comparative study, we conclude the following about the proven effectiveness of the NNS.

Independently administered tests mostly show similar small gains in test scores

Overall gains in maths attainment up to four years after the implementation of the NNS are surprisingly similar across the three independent evaluations. TIMSS test scores improved by some 10 percent in the three to four years following the introduction of the NNS. In comparison, the average point scores on the national standardised tests showed a much smaller increase over the same period: only 3 percent (see Appendix A).

There are indications that at least part of the gains can be attributed to the NNS

None of the studies provide hard evidence of the extent to which the small gain in scores can be attributed to the NNS and later reforms. All studies are based on simple before/after comparisons, which leave the results open to other interpretations. Other national trends may well have played a role. A number of policies and programmes supported the implementation of the NNS. Examples include a prolonged increase in the number of teachers (the average pupil to teacher ratio declined from 17.9 in 1997 to 12.0 in 2008) and establishing the Foundation Stage for children from age 3 to the end of Reception.[21] An indication of wider changes taking place simultaneously with the reform is the similar rise in test scores for science – without a concerted effort to improve science teaching and learning in primary schools like that of the NNS.[22]

The studies provide some indications that the NNS made at least part of the difference, however:

- Effects of the NNS can be traced to relatively strong gains in test scores in areas that were emphasised in the NNS, including the number system and mental calculation
- The international comparative study, TIMSS, shows that gains in test scores over 1995–2003 only show up in year 5 and not in year 9, which suggests that NNS made the difference, since the NNS is aimed at Reception to year 6.

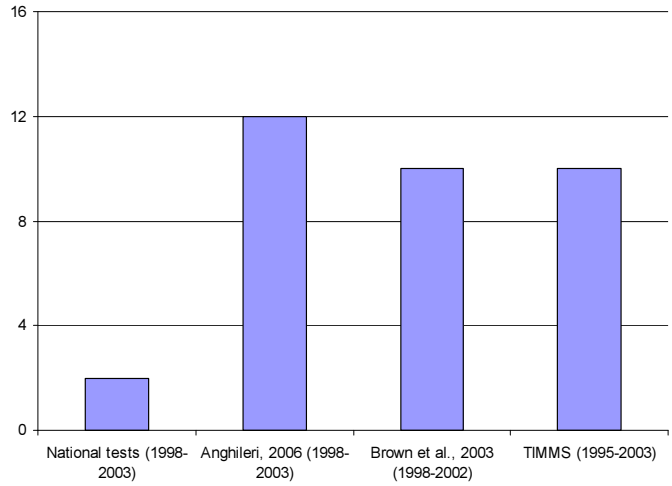
[20] As stated in TIMSS and PIRLS (2004): 'The current assessment includes those topics in mathematics and science that students are likely to have been exposed to up to and including grade 4 and grade 8.'

[21] Reception marks the transition from pre-school to primary education.

[22] It could be that the National Strategies had positive effects on education beyond English and maths. Alternatively, a third factor may have positively affected test scores for all three subjects. See Earl *et al.* (2003: 45) for a discussion.

In addition, classroom practice as it relates to content and lesson structure has evidently changed in line with the NNS and PNS (see below), which is a necessary condition for the reforms to have had an effect.

Figure 4-2: Change in average mathematics test scores after implementation of the NNS



Notes: National tests: average key stage 2 point score for maths (see the Appendix); Anghileri (2006): average year 5 pupil test scores on 8 division problems; Brown *et al.* (2003): average year 4 pupil test scores on number problems; TIMMS: average year 5 pupil test scores. Y-axis denotes the percentage change in test scores compared to the previous period.

Effect of NNS/PNS on classroom practice mostly limited to 'easy-to-implement' changes in curriculum and organisation of lessons; little evidence regardjng more fundamental changes in teaching introduced by the NNS/PNS yet

The reforms were received positively by teachers (Askew *et al.*, 2006: section 3.1) and teacher trainees (Basit, 2003). The more immediately understandable aspects of the NNS seem to have been implemented first (daily maths hour, three-part lesson, change in planning, teaching range of methods). Thus any changes in pupils' attainment that are the result of the NNS are related to the change in format and structure of the lesson. Those reforms that require a deeper understanding of mathematical principles – such as encouraging flexibility in choice of calculation strategy and methods to develop structured written methods – are slow in coming, precluding evaluation. The evaluations find great variation between schools in these higher-order teaching methods.

Interactive whole class teaching is of little benefit to low-attaining pupils, and appears to favour boys over girls

All studies find a greater variance in test scores, with the 25 percent lowest performing pupils showing the smallest increases. The evaluations suggest that the greater variation could be the result of the emphasis on (interactive) whole class teaching. The needs of low-attaining pupils are not fully met in a whole class setting; on average, they made almost no gain in mathematical attainment. Some high-attaining pupils felt frustrated at their progress being held back; on average, their improvement was lower than that of average-performing pupils. Based on interviews and classroom observations, the evaluations suggest that interactive whole class teaching favours boys over girls, with girls perceiving it as

competitive and unpleasant. Just like Earl *et al*. (2003) and Ofsted (2002), the independent evaluations find greater gains in scores for boys than for girls. Only TIMSS does not confirm this result.

CHAPTER 5 Some general observations on evaluation methods

Evaluations (government-initiated and independent) of the impacts of government educational reforms in England have several important limitations, which are important for policy-makers to consider. Mostly, reforms are applied system-wide and are not piloted, which means that it is difficult to conduct randomised control field work. The absence of an understanding of the counterfactual is especially problematic in an area where it is often difficult to attribute impacts to specific factors such as teacher quality, class sizes, pedagogy and types of assessment that influence educational outcomes. In short, it is difficult to understand which aspects of the reform are effective, let alone make judgements on trade-offs between aspects of the reform.

Furthermore, evaluation and monitoring are often not considered at the design stage of the reform. An expert in maths education consulted by this study pointed to the absence of an 'evaluation culture' in the education field in England, which also implies that many education policies are put into place without ex-ante consideration for the evidence base on how effective the interventions are or careful thought on how to conduct ex-post evaluation. Thus, evaluations that track performance over time, follow cohorts of pupils through the system, and systematically assess the factors that could shape this performance are to an extent limited in number. A result is that it is difficult to assess the impact of government reforms in England as few evaluations build strong evidence-based causal or correlational relationships between reforms in the area of numeracy and educational outcomes. Many evaluations look at aspects of implementation, rather than looking at educational outcomes.

Finally, some evaluations solely look at educational outcomes in terms of national test scores, without comparing progress in educational outcomes with independent and international tests on maths ability such as TIMSS (Trends in International Mathematics and Science Study)[23] and PISA (Programme for International Students Assessment)[24] or placing such test scores in the context of wider educational attainment in terms of further study or job outcomes. The former would allow evaluators to understand whether the effect is specific to progress on national scores and understand the reasons behind this, for

[23] See, http://nces.ed.gov/timss/ (accessed June 2008).

[24] See http://www.pisa.oecd.org/pages/0,2987,en_32252351_32235731_1_1_1_1_1,00.html, (accessed June 2008).

instance teaching to the test or problems in the validity and robustness of national testing. The latter would also begin to answer an important aspect of the numeracy strategy: namely whether pupils develop sufficient maths skills in primary education to prepare them for secondary education and eventually the job market.

CHAPTER 6 Additional evidence regarding elements that are central to the reforms

As we have seen in Chapters 3 and 4, the lessons that can be learned from evaluations of the enacted reforms are limited. Evaluating uniformly 'prescribed' reforms like the NNS/PNS appears to have been difficult because of the lack of a natural control group (see also Section 3.7). The before-after research designs left the findings open to alternative interpretations. In addition, the NNS/PNS had been implemented only to a limited extent, precluding evaluation of the 'deeper change' in teaching methods. Finally, the reforms included a package of initiatives, which made it difficult to see which (combination of) measures made the difference and which measures were not effective.

To fill this gap, in this chapter we review additional evidence on changes in primary maths education that are central to the reforms in England, including studies based on experiments across a few schools or evaluations of similar reforms in other countries. The issues on which additional evidence is reviewed are: the effect of school class size on pupil attainment in maths; pedagogy; the effectiveness of whole class teaching versus group or one-to-one teaching; the effectiveness of formative assessment and personalised learning; teacher quality; and the impact of testing on pupil learning. These areas were identified in meetings between the NAO and the RAND Europe study team and were considered especially relevant for evaluating the past and present strategy of the UK Government to improve maths outcomes.[25] This chapter contains short reviews of the main evidence available in these areas. These reviews fall short of a systematic review of the effectiveness of initiatives and do not explicitly judge the research design of referenced research. A word of caution has to be given on evaluations and assessments in the area of education. Effects are often hard to attribute solely to specific interventions and most evaluations tend not to use control groups as interventions are mostly applied across the education system.

6.1 Effect of school class size on pupil attainment in maths

The debate over class sizes has been an integral part of discussions around education reform in the United Kingdom.[26] [27] Indeed, class size reduction has been on the agenda of

[25] See for instance some frequently asked questions on the PNS, http://www.standards.dfes.gov.uk/primary/faqs/#236099 (accessed June 2008).

[26] See for instance recent media attention, http://www.telegraph.co.uk/news/uknews/1563523/OECD-UK-trails-Slovenia-in-class-sizes.html (accessed June 2008).

most Organisation of Economic Cooperation and Development (OECD) countries over the last five years. Every OECD country bar one has increased funding for reducing class sizes and increased its number of teachers (OECD 2005). There has been substantial research on the impact of the reduction of class size on educational outcomes. However, most studies look at the effect of class sizes on general educational outcomes, rather than mathematics outcomes specifically. Though most studies highlight the importance of educational resources in promoting better educational outcomes, most agree that a reduction in class sizes will only have moderate benefit to students. Moreover, a reduction in class size can be a rather expensive policy lever in promoting better educational outcomes.

Hanushek (1989) produced a number of articles and influential literature reviews on the issue of class size reduction in the United States. He analysed over 377 studies on this issue and drew the main findings together in a series of systematic reviews. He found that there is only limited correlation between the level of school inputs and student performance. He measured school performance based on student test scores. There have been a number of qualifications of his work. Hanushek did not examine policy alternatives and only looked at marginal effects. However, even large randomised control studies show a variable effect of class size reductions. The most famous study, at least in the United States, is the Tennessee STAR experiment[28], which found positive effect sizes of around .25 due to class size reductions.[29] Students in the small size classes (13–17 pupils) consistently showed academic gains over their counterparts in the regular size classes (22–25). Statistically significant differences were found on all achievement measures in all subject areas in every year of the experiment. However, California also had a primary school class size reduction and found essentially no effects when controlling for pre-existing differences in the groups (see Stecher et al 2003).

A more recent study by Wiliam (2007) in the United Kingdom looked at the cost-effectiveness of 'formative assessment' and 'reductions in class sizes'. He found that the use of formative assessment is about 20 times more cost-effective than the reduction of class sizes. Cutting class sizes by 30 percent produces an effect in the order of four additional months of educational development per year at a cost of £20,000, whereas an effect of eight months of additional development per year through formative assessment would cost about £2,000 per classroom. So, the Wiliam research shows a positive correlation between class size reduction and educational outcomes, but at a cost.[30] Moreover, Krueger (2000) on the basis of Hanusheks's work identified substantial and significant returns from reducing class sizes in the early grades. He suggests targeting class size reductions to those who would benefit most from it, namely young children and schools in high poverty districts. Similarly, a recent study shows that the impact of class size reductions on low-

[27] See debate on http://www.classsizeresearch.org.uk/ (accessed June 2008).

[28] A summary can be found at http://www.reduceclasssizenow.org/sa_articles/SA11.pdf.

[29]The effect size is the ratio of the average improvements in the test scores of pupils involved in an innovation over the range of scores for typical groups of pupils on these same tests (taken from Black and Wiliam 1998b).

[30] See also Institute of Education website, http://ioewebserver.ioe.ac.uk/ioe/cms/get.asp?cid=1397&1397_1=18281, (accessed May 2007).

attaining students may have been underestimated in the past (Blatchford 2008). The study based on observations of 686 pupils in 27 primary schools and 22 secondary schools showed an 11 percent increase in low-attaining children in primary schools being 'off-task' if class sizes increased by five pupils. This effect in primary schools is less than in secondary schools. The effect was also less pronounced in medium-level and high-attaining pupils. In the past, Blatchford (1998) had also pointed to the possible unintended consequences of smaller class sizes. A negative outcome was that teachers regarded smaller classes as containing more aggressive students. A positive outcome of larger classes not seen in smaller classes was that students were interactive and worked with others more, rather than principally relying on the teacher.

Card and Krueger (1996) also found that test scores might not be the right outcome measure to use. Instead, they looked at educational attainment and earnings. They found a systematic and positive relationship between educational resources and attainment and earnings. In New Zealand, Boozer and Maloney (2001) used longitudinal data to suggest that smaller class sizes were associated with better educational and employment outcomes. However, this was only noted in groups that had benefited from persistent and consistent class size policies.

There has been much debate on the optimal size of classes. Some research in the United States has suggested that positive educational outcomes are associated with classes no greater than 17 pupils (McRobbie and Harman 1998). In the United Kingdom, classes with fewer than 24 pupils have been identified as having a positive impact (Mortimore 1998). Blatchford (2008) quotes other studies that suggest that reductions below 20 or 15 pupils are associated with positive educational outcomes. However, he warns that a sole focus on class size is 'oversimplistic', as it does not take into account other important factors such as teacher quality, type of teaching used, and pupils' engagement.[31]

In summary, the evidence mostly shows that the benefits from reductions in class size are moderate (also distinguishing between educational outcomes and attainment); the effect sizes are variable; and the intervention potentially costly compared to other interventions. If class size reductions are to be part of the policy mix, these reductions should be directed at areas where the evidence shows the largest gains can be made, namely in classes containing low-achieving students and those in the first years of schooling (see also the STAR experiment in Tennessee). Also, class size reductions as a policy lever need, crucially, to be evaluated against other types of interventions, such as for instance formative assessment and interventions aimed at improving teacher quality.

6.2 Choices in pedagogy

Choices in pedagogy can have an impact on educational outcomes. Formative assessment often seen as part of pedagogy is discussed below in some detail. The use of assessment and wider aspects of pedagogy are important parts of the PNS.[32] Similarly, the debate on whole

[31] See also the Institute of Education website, http://www.classsizeresearch.org.uk/publications.html, (accessed May 2007).

[32] See for instance http://www.standards.dfes.gov.uk/primary/about/ (accessed June 2008).

class teaching versus personalised teaching could be categorised as a pedagogic intervention. In terms of other pedagogic interventions, the evaluation evidence is not systematic. There are two main reasons for this. Firstly, the concept is rather nebulous and could encompass a variety of areas and interventions. Secondly, within these areas and interventions, the body of evidence on particular pedagogic interventions is mostly limited. Therefore, this report has focused primarily on 'whole class teaching versus personalised teaching' and 'formative assessment'.

However, there has been some evaluation work on reform or 'constructivist' teaching in maths compared to traditional teaching. The term 'reform-oriented teaching' describes a collection of instructional practices that are designed to engage students as active participants in their own learning and to enhance the development of complex cognitive skills and processes. This represents the main debate in the reform of maths teaching as proposed by the National Science Foundation in the United States. A RAND report by Klein et al. (2000) on teaching in maths and science directly compares two types of practices in schools across the United States that implemented changes in teaching proposed by the National Science Foundation. The findings in the United States provide some support for the hypothesis that reform-oriented teaching is associated with improved student achievement in both maths and science. However, as with most large-scale field studies, there are many factors that may have artificially increased or decreased the observed effect sizes. Nevertheless, the consistency of the results across sites is encouraging.

Another RAND three-year-long study in the United States looked at reform teaching in the area of maths. The study examined the relationship between reform-oriented instruction and student performance in maths and science. At the end of the study, students who had been exposed to more reform-oriented teaching performed better in both maths and science than those who had experienced less, but the differences in scores were small. The relationship between instructional practice and student performance was stronger when performance was measured on items that required problem-solving skills rather than procedural skills, an outcome that is generally consistent with the goals of reform-oriented teaching. These results illustrate the importance of matching performance measures to reform goals in evaluating instructional innovations (Le et al., 2006). However, this study does not offer a direct comparison with traditional teaching.

6.3 Effectiveness of whole class teaching vs. group or one-to-one teaching

Whole class teaching is an important component of the NNS. Some evidence from NAO case studies and other research suggests that pupils dislike being grouped by ability (setting) as it unfairly stigmatises them, although some schools have said that grouping by setting can have a positive effect on achievement, particularly if pupils move in and move out of different sets. A main issue for consideration is also whether grouping by ability disadvantages certain groups. For instance, there is a debate over whether middle-attaining pupils are stretched enough in this approach. In this context, we can look at the debate of whole class teaching versus more personalised teaching and also the use of formative assessment. Again, most studies discuss overall educational outcomes, rather than specifically mathematics outcomes.

In the context of this report, whole class instruction refers to a teacher-led strategy where concepts and materials are taught to the entire class. Small group instruction occurs when students work in groups or pairs. Small group instruction can take many forms, from a teacher-directed agenda to student-initiated dynamics, and from homogeneous groupings (as in the case of ability grouping) to heterogeneous pairings. It can also encompass more cooperative learning structures, where students work collaboratively to solve a problem, or reflect a 'teacher-student' arrangement, where one student serves as an 'expert' for the group.

Multiple reviews have suggested that small group instruction, broadly defined, may have a slight advantage in promoting student achievement relative to whole-class instruction. Meta-analyses conducted by Kulik and Kulik (1987) and by Slavin (1987) found effect sizes of .17 and .34, respectively in favour of small group instruction. In a review of 79 studies that compared student achievement in small group settings and whole class settings, Davidson (1985) found that in nearly 40 percent of the studies, students receiving small group instruction outperformed their whole class settings. Similarly, in a review of 46 studies, Johnson et al (1995) reported that the average student in a small group setting solved more problems correctly than did nearly three-quarters of the control students who worked individually.

Studies that have examined the conditions under which small group instruction is associated with achievement gains have identified several features that promote learning. These include cooperative methods that emphasise both group goals and individual accountability (Grouws and Cebulla 2000; Slavin 1990); a high level of content-related explanations by students (Webb 1991); and manageable group sizes, typically no larger than five members (Johnson and Johnson, 1984). Research about the optimal ability composition of small groups has reported inconsistent findings, with some studies reporting that homogeneous, medium-ability groups demonstrate the greatest benefits (Webb, 1991; Fuchs et al., 1998), while other studies find that homogeneous, high-ability groups show the largest achievement gains (Good, Mulryam et al. 1992).

It is important to keep in mind that the effectiveness of small group (or whole class) instruction frequently depends on unmeasured teacher and student processes that facilitate instruction (Good, Mulryam et al., 1992). Indeed, studies have found that whole-class instruction can be effective, especially when the entire class engages in discussions of their peers' ideas and problem-solving approaches (Wood, 1999). Thus, it may not be the organisational arrangement *per se* that leads to enhanced achievement, but the kinds of teacher and student behaviours manifested in each arrangement that are associated with better outcomes. In general, more research on the processes that promote better learning in both types of arrangement seems warranted.

6.4 Formative assessment

Schools' use of assessment data to tailor the teaching to the pupil's need is often seen as an important component of educational reform, also in the PNS. A related issue is whether it is actually possible for primary schools to identify the gifted, talented children generally and in maths specifically (or do pupils have to be gifted and talented in several areas?) from

the less talented and if so how. A recent initiative by the Government is to move towards the identification of gifted children. This means that given the assumption that a certain percentage of children in a class are gifted or talented, the burden will be on teachers to identify them. Formative assessment is seen as a way to understand students' strengths and weaknesses and indeed improve overall educational outcomes.

Formative assessments, in which teachers use assessment results to provide diagnostic feedback about their students' strengths and weaknesses and guide their instruction, have long been a staple of teachers' instructional practices (Boston, 2002). The most helpful type of feedback on tests and homework provides specific comments about errors and specific suggestions for improvement and encourages students to focus their attention thoughtfully on the task rather than on simply getting the right answer (Bangert-Drowns, Kulik, and Morgan, 1991; Elawar and Corno, 1985). This type of feedback may be particularly helpful to lower-achieving students because it emphasises that students can improve as a result of effort.

Comprehensive reviews of teacher-made formative assessments (e.g., quizzes, tests, and writing assignments) have reported substantial effects on achievement – up to an extra four months of growth per year (Natriello 1987). Black and Wiliam (1998a) conducted a comprehensive review of approximately 250 studies, and found that the effect sizes associated with the use of formative assessments ranged from .5 to 1.0. In a more limited meta-analysis, Black and Wiliam (1998b) on the basis of 20 studies found that the effect size ranged from 0.4 to 0.7. This means, for instance, that an average student benefiting from formative assessment would perform as well as a student in the top 35 percent. They concluded that the effect size of formative assessment tends to be higher than for most other educational interventions. In addition, they concluded that improved formative assessment helps low achievers more than other students. In this way, formative assessment helps reduce the divergence in achievement while raising achievement overall. This is shown in a study (Beaton 1996) devoted to low-achieving students and students with learning disabilities, which shows that frequent assessment feedback helps both groups enhance their learning. Using data from TIMSS, Rodriguez (2004) conducted analysis of teachers' classroom assessment practices and found that greater use of homework and test results for feedback was associated with effect sizes of .5 to 1.8. As noted by Bloom (1984), these effect sizes were as large as the effect sizes found with one-on-one tutoring.

Formative assessment is also shown to be cost-effective against some other interventions proposed in the United Kingdom. A recent study by Wiliam in the United Kingdom compared the cost-effectiveness of formative assessment with reductions in class sizes and found that formative assessment costing £2,000 per classroom was 20 times more effective than class size reduction by 30 percent. A £2,000 investment per classroom would offer eight additional months of educational development per year. [33]

While these mostly United States-based studies underscore the potential of formative assessments, it is important to note that these studies have examined teacher-made formative assessments. Little is known about the use of standardised formative assessments,

[33] See also Institute of Education website, http://ioewebserver.ioe.ac.uk/ioe/cms/get.asp?cid=1397&1397_1=18281, (accessed May 2007).

such as the types put forth by testing companies. An advantage of standardised formative assessments relative to teacher-made tests is that they are designed to explicitly reflect content standards (Stiggins, Frisbie and Griswold, 1989), thereby providing diagnostic information about which specific standards students have and have not attained. This enables teachers to design and revise lesson plans according to student needs.

Despite the intuitive appeal of standardised formative assessments, no empirical studies have yet examined the relationship between teachers' use of these assessments and student achievement. However, there have been some studies that have looked at the factors that facilitate and stymie teachers' implementation of standardised formative assessments (Crooks, 1988; Murnane and Sharkey, 2005; Stiggins, 2001). These include:

- Technology barriers: In order to effectively use data, teachers and administrators need to obtain results quickly. This necessitates a well-organised database and user-friendly software to access the results.

- Teacher knowledge capacity: Studies have shown that few teachers know how to use the feedback from assessment systems to guide their instruction (Cizek, 2000). This suggests that teachers may need specific pre- or in-service training on how to interpret and make use of standardised test results.

- Cultural norms: Teachers are often expected to work collaboratively to analyse the results and implement a plan that addresses grade- and school-level achievement weaknesses. However, teachers may be reluctant to share classroom-level results, as comparison of scores may lead to unfavourable consequences and inappropriate conclusions about teacher effectiveness.

As noted earlier, the literature base on standardised formative assessments is sparse. Indeed, some observers point to the fact that the effect of intervention by formative assessment would be more limited if the assessment were applied across the system. The question remains whether formative assessments when applied across a system remain effective educational innovations. Additional studies that examine the use of such assessments as a means of improving student achievement and their effectiveness as compared to teacher-constructed classroom assessments would further our understanding of strategies that can enhance student learning.

6.5 Quality of teachers

The quality of teachers is an aspect of teaching that has often been overlooked. This is surprising given the evidence of the effect that good teaching can have on educational outcomes. That said, it is difficult to define teaching quality. It can encompass teacher qualifications (such as degree attainment, type of courses taken), professional development (amount and nature of training), as well as frequency of use of teaching practice. Many studies do not necessarily distinguish between the various aspects or attribute effects to specific aspects.

However, in terms of outlining the effect of the quality of teaching, Barber and Mourshed (2007), looking at the world's best education systems, cite 1997 data from Tennessee on

the difference in student outcome for students aged eight starting at the 50th percentile over a three-year period. Those with a high-performing teacher are likely to reach the 93rd percentile while those with a low-performing teacher are more likely to reach the 37th percentile. Other studies (Hedges et al 1994; Darling-Hammond 1998) have also found that teacher quality (qualifications and professional development) has a positive impact on student achievement. Finally, anecdotal evidence in Barber and Mourshed (2007) suggests that countries with the best-qualified and motivated teachers (for instance Finland) have better educational outcomes. So, teacher quality might have a substantial impact on student outcomes over time. This effect might also be more substantial than other interventions, such as reducing class sizes (Barber and Mourshed 2007), though it is difficult to assess this point without adequate randomised control field studies.

It is important to consider that a reduction of class size can have an adverse effect on teacher quality. An increase in the number of teachers could bring in more under-qualified teachers. Other reforms such as reductions in class size can positively (or negatively) affect teaching quality. Here policy-makers need to understand the trade-offs between reducing class sizes and teacher quality, and their potential effects on student outcomes. In addition, policy-makers could look at the most effective ways to support teacher quality, for instance through professional development or through more effective identification of teacher quality in the recruitment process. Such policy would have to take into account evidence regarding what characteristics make a teacher a high-performing teacher (e.g. experience, pedagogy, content knowledge [see Williams Review]) and how teachers can affect student outcomes (see e.g. Hamre and Pianta 2005). However, it is important to note that it is often difficult to identify the main elements of what makes someone a good teacher in terms of attributes and pedagogy and isolate the effect on this basis.

6.6 Impact of testing on pupil learning

The debate over the impact of standardised testing on pupil learning has significant implications for the evaluations that take place. If one agrees with the fact that test scores reflect an adequate understanding of maths skills, then learning how to do exams or indeed 'teaching to the test' might be a perfectly good way to promote better educational outcomes. If one sees a score on a test as an imperfect or even inadequate proxy for a good set of maths skills, then better performance on a test might not be a good indicator of educational achievement. This rather emotive and normative debate raises the issue of teaching to the test and high stakes testing. Because of the high stakes in testing – for teachers, schools, pupils and parents – teaching may become dominated by how to do well on the test rather than how to best understand the underlying principles of maths (Harlen, 2007). If 'teaching to the test' is indeed detrimental to learning (Harlen and Deakin, 2002), then gains in test scores may overestimate actual progress. Therefore, this study includes several evaluations that do not rely on national test score data, but on independently administered tests developed for research purposes.

Independent scores tend to broadly confirm some of the progress noted in the NNS. This is not always the case. A 2001 study of literacy testing in Texas showed a particular divergence between progress noted in state scores compared to more national independent tests, with state scores substantially exceeding national scores and in addition showing a

narrowing between educational outcomes of traditionally white pupils and those from minority groups (Klein et al 2001). This led to various conclusions in the study on the relationship between teaching and how the test was administered:

- Students appeared to be coached for the test at the expense of wider teaching
- The curriculum was narrowed at the expense of specific subjects
- An increase of activities substantially reduced the validity of the scores
- Results were biased by particular features of the testing programme, such as not counting top grades and bottom grades.

Thus, accountability in the testing regime is a key feature.[34] Independent validation and audits of the testing regime seem to be common 'good practice' to monitor the unintended outcomes of high stakes testing, for instance the narrowing of the curriculum. In addition, linking school and teachers' incentives to the outcome of test scores can undermine the validity of the testing regime.

Another often cited impact of high stakes testing is its effect on high achievers. With NNS, there is a reliance on the Key Stage 2 national assessments as 'the indicator' of learning. The target is framed in terms of the percentage of pupils reaching Level 4 on that assessment. Key Stage 2 intervention programmes tend to be directed at the 'not quite Level 4' group, raising the possibility that these children may benefit disproportionately from the intervention efforts. Evidence to date, however, indicates that this has not happened as the entire distribution of scores has moved up (Earl et al., 2003: 46).

Finally, studies do not only look at the impact of standardised testing. More studies are evaluating the impact of classroom evaluation and testing on pupil learning, as discussed in the formative assessment section. Arguably, the effects of classroom evaluations can shape pupil learning more than standardised tests (see Crooks 1998 and Stiggins 2001).

6.7 Conclusion

There are some basic observations about the international evidence. A first observation in most studies is that attributing the effect on educational outcomes to specific factors and controlling for others remains difficult. Secondly, the meanings of concepts in education often overlap or are somewhat nebulous. There are overlaps between the use of formative assessment and teacher quality as a factor in improving educational outcomes and indeed between certain types of pedagogy (types of teaching), formative assessment and teacher quality. Moreover, concepts such as the quality of teachers and pedagogy remain nebulous. Pedagogy can include a wide range of teaching interventions. Quality of teaching consists of aspects such as qualifications, professional development and teaching practice. Therefore, a report like this has to take care not to confuse effects or overemphasise specific effects. Thirdly, the evidence of the impact of many interventions is often not specific to maths skills, but rather focuses on more general educational outcomes, effect sizes, or the

[34] This is also shown in an important recent study on the 'No Child Left Behind' programme by Hamilton et al 2007.

equivalent of additional months of education in a given year. Thus, studies that look at the role of interventions such as reduction of class sizes, formative assessment, and personalised teaching in improving mathematics have to take care that the evidence is specific to the area of maths or indeed generalisable. In general, there is still a need for more and better research that benefits from proper research design such as the use of control groups and control of variables.

That said, the international evidence gives some indications of the effect sizes of specific components of educational reforms, the cost-effectiveness of some components, as well as the trade-offs between some of these components of reform (e.g. the debate on class size reductions and its effect on the quality teaching; the cost trade-offs between investing in class size reductions or formative assessments), which are important to consider when evaluating current UK Government policies. This evidence could also inform future strategies on improving maths skills.

The literature suggests that teacher quality and the use of formative assessment are important factors in improving educational outcomes, while the effect of class size reductions is moderate. Moreover, certain interventions show specific effects on pupil achievement in particular ability groups. Group teaching tends to have the largest effects in medium- and-high ability groups and class size reductions tend to be more effective in low-ability groups and younger-aged pupils. Formative assessment tends to show the largest effects in low-ability pupils. The limited studies available on the impact of high stakes testing show little evidence of the negative impact on maths skills of such testing.

REFERENCES

Reference list

Anghileri, J., 2000, 'Development of division strategies for Year 5 pupils in ten English schools', *British Education Research Journal*, 27 (1), pp. 85–103.

Anghileri, J., 2001, 'Contrasting approaches that challenge tradition', in: Julia Anghileri (*ed.*), *Principles and practices in arithmetic teaching*, Open University Press, Buckingham.

Anghileri, J., 2006, 'A study of the impact of reform on students' written calculation methods after five years' implementation of the National Numeracy Strategy in England', *Oxford Review of Education*, 32 (3), pp. 363–380.

Bangert-Drowns, R. L., Kulick, J. A., and Morgan, M.T. 1991, 'The instructional effect of feedback in test-like events', *Review of Educational Research*, 61 (2): 213–238.

Barber, M. & Mourshed, M. 2007, *How the best performing school systems come out on top*, McKinsey & Company

Basit, T. N., 2003, 'Changing practice through policy: trainee teachers and the National Numeracy Strategy', *Research Papers in Education*, 18 (1), pp. 61–74.

Beaton, A. et al. 1996, *Mathematics Achievement in the Middle School Years*, Boston: Boston College.

Black, P. and Wiliam, D.,1998a, 'Assessment and classroom learning.' *Assessment in Education, 5*, 7–74.

Black, P. and Wiliam, D. 1998b, *Inside the black box: Raising standards through classroom assessment*, Phi Delta Kappan, 80 (2): 139–148. (Available online: http://www.pdkintl.org/kappan/kbla9810.htm)

Blatchford, P and Martin, C, 1998, 'The Effects of Class Size on Classroom Processes: It's a bit like a treadmill – working hard and getting nowhere fast!' *British Journal of Educational Studies*. 46(2),pp 118–137.

Blatchford, P., Russell, A., and Brown, P (in press), 'Teaching in large and small classes.' In L.J. Saha and A.G. Dworkin (Eds) *The New International Handbook of Teachers and Teaching* Springer

Blatchford. P., Russell, A., Bassett, P., Brown, P., and Martin, C. (2006), 'Effects of class size on the teaching of pupils aged 7 to 11 years: implications for classroom management'. AERA Annual Meeting, San Francisco.

Boston, C., 2002, 'The concept of formative assessment.' *Practical Assessment, Research and Evaluation*, 8(9). Retrieved May 9, 2008 from http://PAREonline.net/getvn.asp?v=8&n=9

Brown, M., Askew, M., Baker, D., Denvir, H. and Millett, A. , 1998, 'Is the National Numeracy Strategy research-based?', *British Journal of Educational Studies*, 46 (4), pp. 362–385.

Brown, M., Askew M., and Millett A., 2003, 'How has the National Numeracy Strategy affected attainment and teaching in year 4?', in: Williams, J. (Ed.) *Proceedings of the British Society for Research into Learning Mathematics*, 23(2), pp. 13–18.

Brown, M., Askew M., Rhodes V., Denvir H., and Ranson E., 2006, 'Evaluating the effects of pedagogy on numeracy learning', chapter 3 in: Mike Askew, Margaret Brown and Alison Millett (eds.), *Learning about number: interactions and outcomes in primary classrooms*, Springer.

Boozer, M. A. and Maloney, T., 2001, 'The Effects of Class Size on the Long-run Growth in Reading Abilities and Early Adult Outcomes in the Christchurch Health and Development Study', Working Paper 01/14, The Treasury, Wellington.

Card, D and Krueger, A, 1996, 'School Resources and Student Outcomes: An Overview of the Literature and New Evidence from North and South Carolina', NBER Working paper 5708.

Cizek, G. J. (2000). 'Pockets of resistance in the assessment revolution.' *Educational Measurement*: *Issues and Practice*, *19*, 16–23, 33.

Crooks, T. J. (1988). 'The impact of classroom evaluation practices on students.' *Review of Educational Research*, *58*, 438–481.

Darling-Hammond, L 1998, 'Teachers and Teaching: Testing Policy Hypotheses from a National Commission Report', *Educational Researcher*, 27(1), pp 5–15.

DfEE, 1999, The National Numeracy Strategy. Framework for teaching mathematics from Reception to Year 6, London.

DfES, 2003, Excellence and Enjoyment, London.

Earl, L., Fullan M., Leithwoord K., and Watson N., 2000, 'Watching and Learning', OISE/UT Evaluation of the National Literacy and Numeracy Strategies, Department for Education and Skills, London.

Earl, L., Levin B., Leithwood K., Fullan M., and Watson N., 2001, 'Watching and Learning 2', OISE/UT Evaluation of the National Literacy and Numeracy Strategies, Department for Education and Skills, London.

Earl, L., Watson N., Levin B., Leithwood K., Fullan M., Torrance N., Jantzi D., Mascall B., and Volante L., 2003, 'Watching and Learning 3.' Final report of the OISE/UT evaluation of England's national literacy and numeracy strategies, Department for Education and Skills, London.

Elawar, M. C., and Corno, L. 1985, 'A factorial experiment in teachers' written feedback on student homework: Changing teacher behaviour a little rather than a lot.' *Journal of Educational Psychology*, 77 (2): 162–173.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., and Karns, K. (1998). 'High-Achieving Students' Interactions and Performance on Complex Mathematical Tasks as a Function of Homogeneous and Heterogeneous Pairings.' *American Educational Research Journal,* 35(2): 227–267.

Gibbons, S., and Machin S., 2003, 'Valuing English Primary Schools', *Journal of Urban Economics*, 53 (2), pp. 197–219.

Good, T. L., Mulryan C., and McCaslin M. (1992), 'Grouping for Instruction in Mathematics: A Call for Programmatic Research on Small-Group Processes' In Grouws, D. A. (ed.) *Handbook of Research on Mathematics Teaching and Learning.* New York, MacMillan: 165–196.

Grouws, D. A., and Cebulla, K. J. (2000), *Improving student achievement in mathematics: Recommendations for the classroom* (Educational Practice Series–4). Brussels, Belgium: International Academy of Education. (ERIC Document Reproduction Service No. ED463953).

Hamre, B. K., and Pianta, R. C., 2005, 'Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure?', *Child Development* 76 (5) (2005), pp. 949–967.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., Sloan McCombs J., Robyn, A., Lin Russell, J., Naftel, S., and Barney H., 2007, *Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States*, RAND: Santa Monica MG 589.

Hanushek, E. A. (1986) The Economics of Schooling: Production and Efficiency in Public Schools, *Journal of Economic Literature* 24 (September), pp 1141–1177.

Hanushek, E. A. (1989) 'Expenditures, Efficiency, and Equity in Education: The Federal Government's Role', *American Economic Review* 79(2), pp 46–51.

Harlen, W. and Deakin Crick, R., 2002, *A systematic review of the impact of summative assessment and tests on students' motivation for learning*, Research Evidence in Education Library, EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen, W., 2007, 'The Quality of Learning: assessment alternatives for primary education', Primary Review Research Survey 3/4, University of Cambridge Faculty of Education, Cambridge.

Harris, S., Keys, W. and Fernandes, C., 1997, *Third International Mathematics and Science Study, Second National report, Part One, Achievement in Mathematics and Science at Age 9 in England*, Slough, National Foundation for Educational Research.

Hedges, L, Laine, R and Greenwald, R. 1994, 'Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes', *Educational Researcher*, 23 (3), pp 5-14.

Johnson, D. and. Johnson, R. (1984). *Circles of Learning*, Washington, DC: Association for Supervision and Curriculum Development.

Johnson, D., Johnson, R. and Johnson, E. (1995). *The New circles of learning: Cooperation in the classroom and school.* Edina, MN: Interaction Book Company.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., Stecher, 2001, *What Do Test Scores in Texas Tell Us?*, RAND Corporation: Santa Monica, http://www.rand.org/pubs/issue_papers/IP202/.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Robyn, A., Burroughs, D., 2001

Krueger, A. 2000, 'Economic Considerations and Class Size', Working Paper #447, Industrial Relations Section, Princeton University, www.irs.princeton.edu/pubs/pdfs/447.pdf.

Kulik, J. A., and Kulik, C.-L. C. (1987). 'Effects of ability grouping on student achievement'. *Equity and Excellence*, *23*, 22–30.

Kyriacou, C., 2005, 'The impact of daily mathematics lessons in England on pupil confidence and competence in early mathematics: a systematic review', *British Journal of Educational Studies*, 53 (2), pp. 168–186.

Le, V., Stecher, B, Lockwood, J. R. , Hamilton, L. S., Robyn, A., Williams, V. L., Ryan, G., Kerr, K., Martinez, J. L., Klein, S. P., 2006, *Improving Mathematics and Science Education: A Longitudinal Investigation of the Relationship Between Reform-Oriented Instruction and Student Achievement*, RAND Corporation: Santa Monica, http://www.rand.org/pubs/monographs/MG480/.

Machin, S., McIntosh, S., Vignoles, A. and Viitanen, T., 2001, *Basic skills, soft skills and labour market outcomes: secondary analysis of the National Child Development Study*, Research Report 25, Centre for the Economics of Education, Department for Education and Skills, London.

McRobbie J, Finn J. and Harman P, 1998, 'Class Size Reduction: Lessons Learned from Experience', WestEd PolicyBrief 23, www.wested.org/cs/wew/view/rs/119.

Mortimore, P. 1998, *The Road to Improvement: Reflections on School Effectiveness*, Swets and Zeitlinger Publishers, The Netherlands.

Muijs, D. and Reynolds, D. 2000, 'School effectiveness and teacher effectiveness in mathematics: some preliminary findings from the evaluation of the Mathematics Enhancement Programme (Primary)', *School Effectiveness and School Improvement*, 11 (3), pp. 273–303.

Murnane, R. J., Sharkey, N. S., and Boudett, K. P. (2005). 'Using student assessment results to improve instruction: Lessons from a workshop.' *Journal of Education for Students Placed at Risk*, *10*(3), 269–280.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*(2), 155–175.

OECD 2005, *Attracting, Developing , and Retaining Effective Teachers*, Paris.

Rodriguez, M. C. (2004). The Role of Classroom Assessment in Student Performance on TIMSS. *Applied Measurement in Education*, 17 (1), 1–24.

Ruddock, G., Sturman, L., Schagen, I., Styles, B., Gnaldi, M. and Vappula, H., 2004, *Where England Stands in the Trends in International Mathematics and Science Study (TIMSS) 2003*. National Report for England, National Foundation for Educational Research.

Slavin, R. E. 1987, 'Ability grouping and student achievement in elementary schools: A best-evidence synthesis.' *Review of Educational Research*, 57, 293–336.

Slavin, R. E. 1990, 'Student team learning in mathematics.' In Davidson, N. (Ed.), *Cooperative learning in math: A handbook for teachers*. Boston: Allyn and Bacon, (pp. 69-102).

Stecher, B. M., McCaffrey, D. F. and Bugliari, D. 2003, The relationship between exposure to class size reduction and student achievement in California, *Education Policy Analysis Archives*, 11(40).

Stiggins, R. J. 2001, The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practice*, *20*(3), PP 5–15.

Stiggins, R. J., Frisbie, D. A., and Griswold, P. A. 1989, 'Inside high school grading practices: Building a research agenda.' *Educational Measurement: Issues and Practice*, *8* (2), 5–14.

Tymms, P., 2004, 'Are standards rising in English primary schools?', *British Educational Research Journal*, 30 (4), pp. 477–494.

Webb, N. M. 1991, 'Task-Related Verbal Interaction and Mathematics Learning in Small Groups.' *Journal for Research in Mathematics Education, 22*(2): 366–389.

Wiliam, D., and Lester Jr, F. K. 2008, 'On the purpose of mathematics education research: making productive contributions to policy and practice.' In English, L. D. (Ed.), *Handbook of international research in mathematics education* (2nd ed., pp. 32–48). New York, NY: Routledge/Taylor and Francis.

Wood, T. 1999, 'Creating a context for argument in mathematics class.' *Journal for Research in Mathematics Education*, 30, 171–91.

APPENDICES

Appendix A: Methodology

To compare the results of independently administered tests with the results of the national standardised tests, average point scores are used. In this appendix, we present the method by which we computed average points scores on the Key Stage 2 test for mathematics.

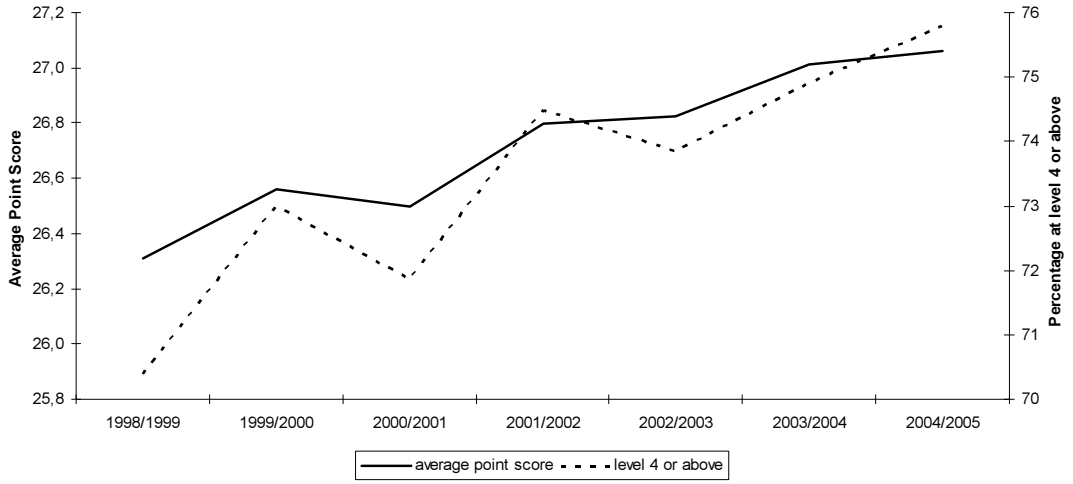
Source of data is the National Pupil Database. For each individual pupil, the result of the Key Stage 2 test is expressed as the level attained, with levels varying from 1 to 6. In line with the value added calculations of the Department for Children, Schools and Families, each level of attainment has a certain score. The scores are as follows:

- Pupils who were working below the level of the test, pupils who took the test but failed to register a level, and Level 2: 15 points
- Level 3: 21 points
- Level 4: 27 points
- Level 5: 33 points

The first group constitutes some one in twenty pupils; one in six achieved level 3; almost half of pupils, level 4; one-third of pupils, level 5.

Based on this scoring method, we computed the average point score for every year. The results are presented in Figure A.1 below. Clearly, the pattern is very similar to the percentage of pupils at level 4 or above. The size of the change is much smaller when based on average point scores rather than the percentage above the threshold, however.

Figure A.1 Average point score closely follows percentage level 4 or above on Key Stage 2 test for mathematics



Source: National Pupil Database